Transcriptome-based predictive modeling approaches in *Arabidopsis thaliana*

Rahul Bhosale

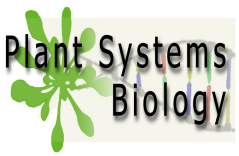
Promoter: Prof. Dr. Ir. Steven Maere
Co-Promoter: Prof. Dr. Lieven De Veylder

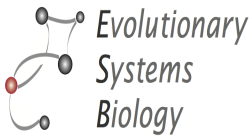
Ghent University (UGent) / Flanders Institute for Biotechnology (VIB)
Faculty of Sciences
UGent Department of Plant Biotechnology and Bioinformatics
VIB Department of Plant Systems Biology

This research was funded by the Scientific Research Flanders (FWO-Vlaanderen) project grant number : G002911N.

Dissertation submitted in fulfilment of the requirements for the degree of Doctor (PhD) in Sciences, Biochemistry and Biotechnology.

Academic year: 2014-2015





Examination Committee

Prof. Dr. Geert De Jaeger (chair)

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Prof. Dr. Ir. Steven Maere (promoter)

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Prof. Dr. Lieven De Veylder (co-promoter)

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Prof. Dr. Malcolm J Bennett*

Faculty of Plant Sciences, School of Biosciences, University of Nottingham

Prof. Dr. Ir. Gerrit T.S. Beemster*

Laboratory for Molecular Plant Physiology and Biotechnology, Department of Biology, University of Antwerpen

Prof. Dr. Alain Goosens*

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Prof. Dr. Lennart Martens*

Faculty of Medicine and Health Sciences, Department of Biochemistry, Ghent University

** Member of the reading commission*

Acknowledgements

"If I have seen further than others, it is by standing upon the shoulders of giants."
-Sir Issac Newton

My promoters, collaborators, colleagues, friends and family have contributed immensely to the success of my PhD studies at VIB, University of Ghent. First of all, I would like to express my deepest gratitude to Professor Steven Maere and Professor Lieven De Veylder for showing confidence in me and giving me the opportunity to work on several multi-disciplinary projects. Their utmost patience and encouragement has helped me to ultimately successfully deliver these challenging projects with confidence and zeal. They have also given me the freedom and numerous excellent opportunities to collaborate and discuss with experts in the field and present my research at major conferences. This has truly helped me to develop my scientific acumen and expertise in the field of Plant Systems Biology and hone my professional skills as a researcher.

My doctoral study has become a reality owing to the generous financial support for fours years from the Dehousse Fellowship, University of Ghent and FWO research grants of Professor Steven Maere and Professor Lieven De Veylder, to which I am sincerely grateful. I would also like to thank Professor Dirk Inze and other Management Committee members at VIB, Department of Plant Systems Biology, University of Ghent, for providing me with the PhD extension grant for six months in order to finish my doctoral studies.

My successful and smooth PhD experience has also been a culmination of the timely support and advice provided to me by my colleagues and the administrative and IT staff. I would like to thank Dr. Veronique Boudolf, Gert Van Isterdael, Fabiola Cuevas, Ilse Vercauteren, and other members of Cell Cycle group for their constant help in conducting experiments. I would like to thank Kevin Vanneste, Jayson Gutierrez, Brigida Gallone, other members of Evolutionary System Biology group for their continuous help and support in resolving difficulties in computational work. I would also like to thank Diane Hermie, Christine Tire, Sophie Maebe, Nathalie Vanden Hautte, Delphine Verspeel, Bernard Vanassche, Christa Verplancke and all other members of the Department of Plant Systems Biology for their ancillary support in matters of administration, accommodation, travels and finances. Nonetheless, I would like to thank Luc Van Wiemeersch, Frederik Delaere and other members of IT help-desk for their constant support in resolving cluster and storage issues during the last four years.

I am also deeply thankful to my family back home for their constant love and support. My parents and my sister have provided an encouraging upbringing to me with the freedom to choose my interests and career options. They have taught me to be sincere and dedicated to my work. I also would like thank Aditi Borkar for being my pillar of support, encouragement and strength through all the testing times and for similarly keeping me grounded and focussed during the euphoric phases in the last four years. Aditi has motivated me to be confident and practical in life and thus helped me to grow both personally and professionally. I consider myself extremely fortunate to have found my best friend in her. I am very thankful to god for blessing me with such a wonderful family. Last, but not in anyway least, I also highly appreciate the support extended by my friends Kevin Vanneste, Stephane Rombauts, Jens Hollunder, Jayson Gutierrez and all others here for making my experience at Ghent a really brilliant one. It felt home away from home only because of them.

Contents

List of Figures

List of Tables

List of Abbreviations

2D two-dimensional

A Adenine

ABA Abscisic Acid

ABRC *Arabidopsis* Biological Resource Center

ANOVA Analysis of Variance

APC/C Anaphase-Promoting Complex/Cyclosome

APK1 APS Kinase 1

ARR *Arabidopsis* Response Regulators

AZF1 *Arabidopsis* Zinc-Finger Protein 1

BHLH Basic Helix-Loop-Helix

BiNGO Biological Networks Gene Ontology tool

C Cytosine

CaCl2 Calcium Chloride

CCS52A2 Cell Cycle Switch protein 52 A2

CDC Cell Division Cycle

CDF Computable Document Format

CDK Cyclin-Dependent Kinase

CDKA A-type CDK

CDKB B-type CDK

cDNA Complementary DNA

CDT Cytolethal Distending Toxin

CKI Cyclin-dependent Kinase Inhibitor

COP1 Constitutively Photomorphogenic

COR Cortex

CORNET Correlation Network

CPC Caprice

CPR5 Constitutive Expression Of Pathogenesis-Related Genes 5

CRT3 Calreticulin 3

crys cryptochromes

CTR1 Constitutive Triple Response 1

CYC Cyclin

CYCA A-Type CYC

CYCB B-Type CYC

CYCD D-Type CYC

CYCH H-Type CYC

DAPI 4',6-diamidino-2-phenylindole

DEL DP-E2F-LIKE

df degrees of freedom

DMSO Dimethyl Sulfoxide

DNA Deoxyribonucleic Acid

DP Dimerisation partner

E2F adenovirus E2 promoter binding Factor

EGL3 Enhancer of Glabra 3

EI Endoreplication Index

EIL1 EIN3-like 1 protein

EIN3 Ethylene-Insensitive 3

ENIGMA Expression Network Inference and Global Module Analysis

eQTL expression quantitative-trait-locus

FACS Fluorescence-activated cell sorting

FDR False Discovery Rate

FWER Family Wise Error Rate

G Guanine

G1 Gap1 Phase

G2 Gap2 Phase

GEO Gene Expression Omnibus

GFP Green Fluorescence Protein

GL Glabrous

GO Gene Ontology

GTL1 GT2 Like 1

GUS β-glucuronidase

HCl Hydrogen Chloride

HECT Homologous to the E6AP Carboxyl Terminus

HFR1 Long Hypocotyl in Far-Red1

HPC High-Performance Computing

HPLC High Performance Liquid Chromatography

HPY High Ploidy

HY5 Elongated Hypocotyl 5

IAA Indole-3-acetic acid

IAR3 IAA-Alanine Resistant 3

ICK3 Interactor of CDC2 kinase

IEA Inferred from Electronic Annotation

ILL6 IAA-amino acid hydrolase ILR1-like 6

ILP1 Increased Level of Polyploid1-1D

ILR IAA-Leucine Resistant

IPD1 Increased Polyploidy Level in Darkness 1

ISS Inferred from Sequence or structural Similarity

JA Jasmonic Acid

JA-Ile Jasmonyl-Isoleucie

JAZ1 Jasmonate-ZIM-Domain Protein 1

KCl Potassium Chloride

KRP Kip-Related Protein

LAF1 Long After Far-red Light1

LIMMA Linear Models for Microarray data

M Mitotic Phase

MCM Mini Chromosome Maintenance

MCSA Monte Carlo-Simulated Annealing

meJA Methyl Jasmonate

MES 2-(N-morpholino)ethanesulfonic acid

MgCl2 Magnesium Chloride

MID Midget

MM Miss Match

MPK1 Mitogen-Activated Protein Kinase 1

mRNA messenger-RNA

MS Murashige and Skoog

MSA M-phase-Specific Activator

MYB R MYB Repeats

NaCl Sodium Chloride

nLogL negative Log Likelihood

ORC Origin Recognition Complex

PCC Pearson Correlation Coefficient

PCNA Proliferating Cell Nuclear Antigen

phys phytochromes

PM Perfect Match

RBR RetinoBlastoma Related

RCA Reviewed Computational Analysis

RHL Root Hairless

RMA Robust Multi-array Average

RNA Ribonucleic Acid

RNA-Seq RNA Sequencing

RNAP RNA Polymerase

RNR Ribo-Nucleotide Reductase

rpm revolution per minute

rRNA Ribosomal RNA

RSL4 Root Hair Defective 6-Like 4

RT-PCR Reverse Transcription Polymerase Chain Reaction

S DNA Synthesis Phase

SAGE Serial Analysis of Gene Expression

SCM Scrambled

SCR Scarecrow

SMR Siamese Related

SSQ sum of squared errors

ST Spatio-Temporal

T Thymine

T-DNA Transfer DNA

TAIR The *Arabidopsis* Information Resource Release

TGA5 TGACG Motif-Binding Factor 5

TOPOVI TOPOISOMERASE VI

tRNA Transfer RNA

TRY Triptychon

TTG1 Transparent Testa Glabra 1

U Uracil

UPL3 Ubiquitin-Protein Ligase 3

UV UltraViolet

WER Werwolf

WOL Wooden Leg

YFP Yellow Fluorescence Protein

Research goals

One of the major challenges in plant systems biology is to understand how plants respond to various environmental signals at their different levels of organisation i.e. cell type, tissue, organ and organism. These signals e.g. water levels, temperature, etc. can be either macro- or micro-environmental in nature. Plants transduce these signals and use (i) transcriptional mechanisms to appropriately control stress-related gene expression levels, (ii) post-transcriptional mechanisms (based on alternative splicing, RNA processing as well as RNA silencing) to define the actual transcriptome supporting the stress response and (iii) post-translational modifications (such as protein phosphorylation, ubiquitination and sumoylation) to regulate the activation of pre-existing molecules to ensure a prompt response to stress. These mechanisms ultimately allow plants to modulate their development and physiology to cope with stress conditions. This thesis focuses only on stress responses at the transcriptional level. The advance in gene expression profiling technologies enables us to measure change in gene expression at any level of organisation in a high-throughput manner. However, such multidimensional data can not be interpreted as is, because of their complex nature. Thus, predictive modeling approaches for interpreting such data have become common practice. Predictive modeling approaches generally estimate the probability of an outcome given a particular input. Such approaches (or models) allow for formulating testable hypotheses, which can be experimentally verified. In many cases, models might provide insight into questions that are hard to address experimentally. In this PhD thesis, we used predictive modeling approaches based on gene expression data for the model organism *Arabidopsis thaliana* to address two major challenges in current plant systems biology, as summarised below.

1. Prediction of gene functions from expression variation due to subtle micro-environmental perturbations across individual wild-type *Arabidopsis* plants

Standard network-guided gene function prediction approaches use datasets produced using a traditional expression profiling setup. Plants are usually grown under a tightly controlled experimental setup, subjected to a single treatment that is usually rather harsh, in order to mask the unwanted residual effects, and pooled to suppress variability among individuals and increase the experimental reproducibility. However, even after taking such precautions, the reproducibility of expression profiling experiments is often poor[3]. Moreover, harsh single treatments are often unrealistic in a natural context, because individual plants in the field are generally simultaneously exposed to multiple subtle changes in the environmental

conditions. Additionally, harsh perturbations force the system to work away from homeostasis to counteract their effects, and may thus lead to off-target responses that are only indirectly related to the perturbations applied. Moreover, it is often practically infeasible to define and perform the hundreds of controlled perturbations needed to unravel particular plant process. To circumvent these problems, we aimed to assess the information contained in the expression variation among individual wild-type plants arising due to subtle uncontrolled perturbations in the growth conditions and its use for reverse engineering purposes. We reanalysed the individual plant gene expression data set generated by Massonnet and co-workers[4] and compared its functional prediction performance with that of same-sized compendia of *Arabidopsis* gene expression experiments, profiling the response to controlled harsh treatments on pooled plant samples (Chapter 2).

2. Prediction of endoploidy distributions in cell type, tissue and organ under developmental and environmental cues

Endoreplication or the endocycle is a variant of the mitotic cell cycle during which cells duplicate their genome (repeatedly) without mitosis, thus resulting into increased nuclear content i.e. endoploidy. Endoreplication is often seen in plants as a prominent response to stress conditions such as DNA damage, Ultra Violet (UV) radiation stress, pathogen attack, etc. Although over the recent years many genes have been identified that control endoreplication onset and progression, lack of a clear knowledge on the temporal and spatial occurrence of endoploidy distribution in an endoreplicating species has hampered the possibility to study the physiological role of the endocycle at the plant level. A major open question is how cells of multiple cell types or tissues with different DNA endoploidy levels are integrated into a developing organ, and how this organisation contributes to the growth of the plant under different environmental conditions. We aimed to use a predictive modeling approach based on gene expression data to obtain a spatio-temporal (ST) endoploidy map of the developing *Arabidopsis thaliana* root and endoploidy map changes in response to various environmental conditions (Chapter 4).

Chapter 1

Introduction to gene function prediction

“If we knew what it was we were doing, it would not be called research, would it?”

Albert Einstein.

For the author contributions, see page 18.

1.1 Basics of molecular biology

1.1.1 Genes : working subunits of DNA

Deoxyribonucleic acid (DNA) is an information storage molecule, which contains all of the instructions a cell requires to sustain itself. These instructions are found within genes, which are sections of DNA made up of specific sequences of four different nitrogenous bases (nucleobases): adenine (A), cytosine (C), guanine (G), and thymine (T). These nucleobases are positioned sequentially on two long anti-parallel polymer strands in a double helix configuration where hydrogen bonds link complementary bases on opposite strands: A with T and C with G[5]. Inside the nucleus, DNA doesn't appear in the naked state but forms a complex with nuclear proteins called histones. The DNA-histone protein complex is called chromatin and is formed when the double helix DNA wraps around histone in a beads-on-a-string configuration and then wraps around itself multiple times to condense into a smaller volume. Such condensed chromatin molecules, i.e. chromosomes, often appear as X-shaped structures inside the cell's nucleus during the cell division process. A concise overview of DNA inside the nucleus of a cell is presented in Figure 1.1. Genes code either for a messenger-ribonucleic acid (mRNA) encoding the amino acid sequence in a polypeptide chain or for a functional RNA molecule. DNA thus holds the information to build and maintain an organism's cells and pass genetic traits to offspring.

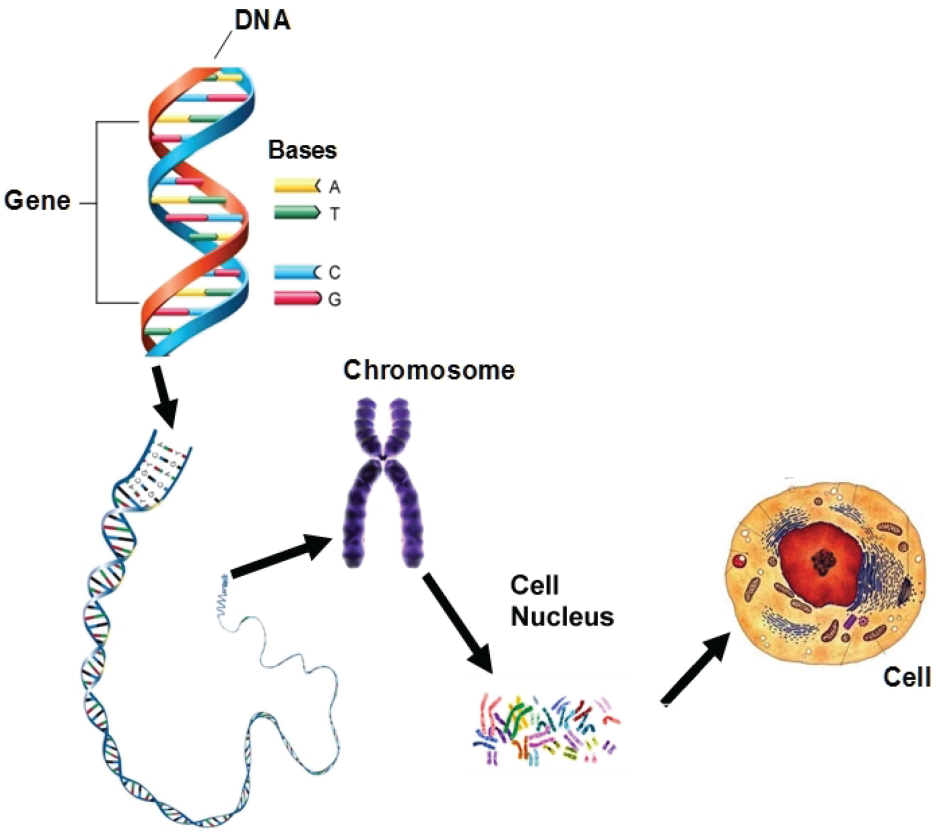


Figure 1.1: An overview of DNA inside the nucleus of cell. Genes are the segments of DNA and are made up of specific sequences of four different nucleobases A,T,G and C and contain instructions to make functional products and maintain the organism's cell. During cell division process, DNA molecules are systematically packed into a number of chromosomes and appear as X-shaped structures inside the cell's nucleus. Picture is taken from wikipedia commons.

1.1.2 Gene expression : two step process

The instructions stored within protein-coding genes are read and processed by a cell for synthesis of a functional gene product in two steps : transcription and translation, the process is collectively termed gene expression. An overview of the flow of information from gene (DNA) to functional gene product (protein)[6,7] in a eukaryote is represented in the Figure 1.2. During transcription, RNA polymerase (RNAP) uses a portion of the cell's DNA as a template and synthesises a strand of RNA that is complementary to the DNA template strand. RNA is chemically similar to DNA, except for three main differences (i) in RNA, a base called uracil (U) replaces T as the complementary nucleotide to A, (ii) RNA is made in a single-stranded, non-helical form while DNA is almost always in a double-stranded helical form and (iii) RNA contains ribose sugar molecules, which are slightly different than the deoxyribose molecules found in DNA. In some cases, the newly created RNA molecule undergoes cleavage at the both ends to generate a finished product such as transfer RNA (tRNA) or ribosomal RNA (rRNA), that serves an important function within the cell. In other cases, the newly synthesised RNA molecule further undergoes extensive changes resulting in mRNA that carries the DNA's message to the ribosome in the cytoplasm where proteins are synthesised. These changes involves splicing, in which noncoding nucleotide sequences, called introns, are clipped out of the mRNA strand, and coding nucleotide sequences, called exons, are retained. Then, a sequence of adenine nucleotides called a poly-A tail is added to the 3' end of the mRNA molecule. This sequence signals to the cell that the mRNA molecule is ready to leave the nucleus and enter the cytoplasm, where that molecule can be translated into protein.

Translation is the process by which a protein is synthesised from the instructions contained in an mRNA molecule. During this process, an mRNA sequence is read using the genetic code, which is a set of rules that defines how an mRNA sequence is to be translated into the 20-letter code of amino acids, the building blocks of proteins. The genetic code is a set of three-letter combinations of nucleotides called codons, each of which corresponds with a specific amino acid or stop signal. This process occurs in a structure called the ribosome, which has a small and a large subunit and is a complex molecule composed of several ribosomal RNA molecules and a number of proteins. Translation occurs in three stages: initiation, elongation, and termination. During initiation, the small ribosomal subunit binds to the start of the mRNA sequence. Then a tRNA molecule carrying the amino acid methionine binds to what is called the start codon of the mRNA sequence. The start codon in all mRNA molecules has the sequence AUG and codes for methionine. Next, the large ribosomal subunit binds to form the complete initiation complex. During the elongation stage, the ribosome continues to translate each codon in turn. Each corresponding amino acid is added to the growing chain and linked via a peptide bond. Elongation continues until all of the codons are read. Lastly, termination occurs when the ribosome reaches a stop codon (UAA, UAG, and UGA). The new protein is then released, and the translation complex comes apart.

Functional proteins are responsible for a wide variety of functionalities in the cell, ranging from enzymatic activity to cellular signalling and structural roles[8]. The rate of production of functional proteins in the cell is regulated at many stages of gene expression, primarily at the level of transcription but also at post-transcriptional, translational and post-translational levels. The diversity of cell phenotypes which are produced from identical genomes is primarily due to differences in gene expression, whether between

different cell types in a multicellular organism, or as a result of diverse gene expression responses between different physiological conditions or developmental stages.

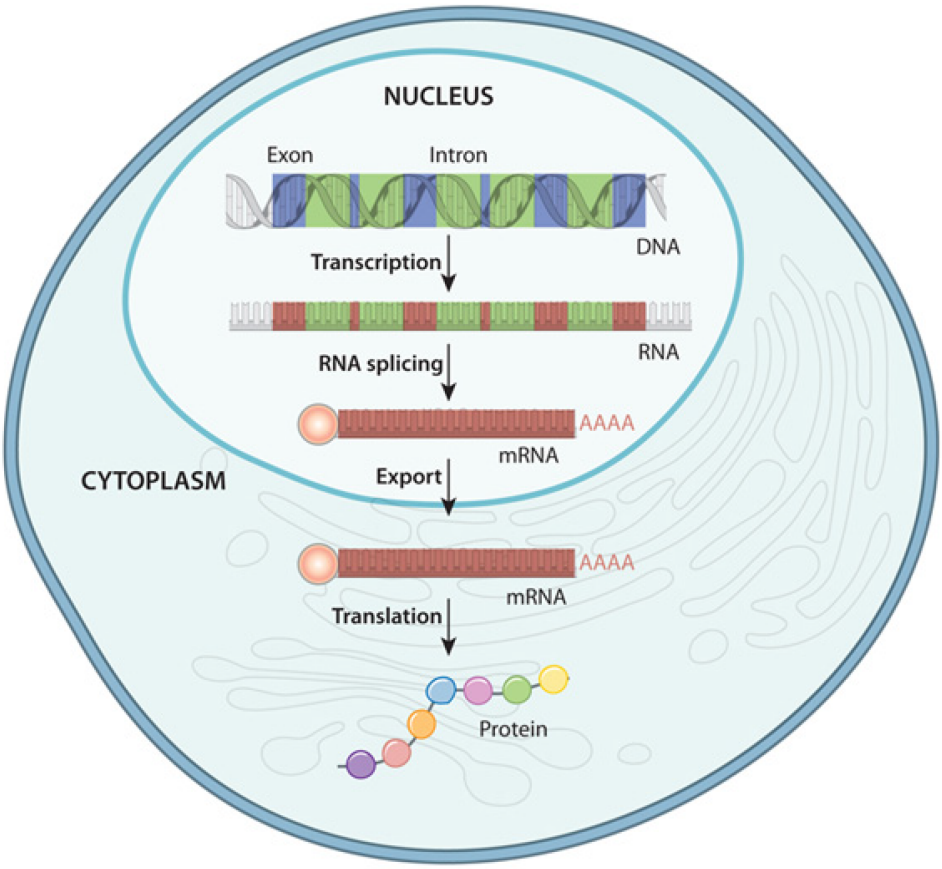


Figure 1.2: An overview of the flow of information from DNA to protein in a eukaryote. First, both coding and noncoding regions of DNA are transcribed into mRNA. Non coding regions (introns) are removed during initial mRNA processing and the remaining coding regions (exons) are then spliced together. The spliced mRNA molecule (brown) is then prepared for export out of the nucleus through addition of an end cap (sphere) and a polyA tail. Once in the cytoplasm, the mRNA can be used to synthesise a protein. Picture is taken from Scitable (`http://www.nature.com/scitable`).

1.2 Emergence of systems biology

The discovery of DNA structure in 1953 by Watson and Crick[5] laid the foundation for studying vital cellular processes such as replication, transcription and translation in molecular terms, essentially the field of molecular biology. Until the past decade, research in molecular biology has been mainly based on a reductionistic view, the idea that complex system can be understood by the analysis of their simpler individual components, i.e. that the intricate operating system of living organisms can be understood by investigating individual genes or proteins at a time. As reductionistic methodologies reduce the number of experimental variables and facilitate analyses, they have been extensively used in the latter half of the 20th century. Nevertheless, these methodologies have limitations. For instance, experimental observations obtained *in vitro* with isolated components of cells are not directly applicable to the physiology of whole organisms. In the past decade, the advances of genome sequencing and high-throughput functional genomics technologies gave rise to the field of systems biology, which contrasts the reductionistic view as

it investigates the behaviour and relationships of all of the components in a particular biological system while it is functioning.

Systems biology approaches can be of two types: (1) top-down, starting from omics-scale data and seeking to unravel the underlying explanatory principles or (2) bottom-up, starting with properties of single molecules and deriving large-scale models that can subsequently be tested and validated. The first approach involves the systematic perturbation (genetical, chemical or environmental) of a biological system; measuring the gene, protein or informational pathway responses; integration of these data and finally formulating mathematical models that describe the structure of the system and its response to individual perturbations[?]. The second approach is more mechanistic but it similarly produces models of a system's behaviour in response to perturbation that can be tested experimentally. In recent years, the increased emphasis on pathways, networks, and systems has given rise to powerful new experimental and bioinformatics methods. Although genomic[9], gene expression[10], and proteomics[11] analyses are now becoming ordinary in biology, the construction of synthetic regulatory circuits[12], the modeling of complex genetic[13] and metabolic networks[14], the measurement of transcriptional dynamics[15] and gene regulation in single cells[16] are just some of the new techniques gaining scientific interest for analysing the properties of complex systems.

Below, section 1.3 describes the experimental techniques used to generate transcriptome data and section 1.4 describes the computational methods that are used to analyse and integrate huge amounts of transcriptome data, as used in this thesis. Finally, some relevant models for predicting gene functions by perturbing the biological system are described in section 1.5.

1.3 Generating transcriptome data

1.3.1 High-throughput technologies

Gene expression is a dynamic process, as the same gene may be turned on or off in a particular cell under different conditions or it may be expressed in different quantities. The abundance of a transcript under a particular condition reflects a dynamic balance between production (transcription) and degradation of the mRNA concerned. In the past decades, gene expression was studied by looking at only one or a very few genes at once using a method called the Northern blot[17]. For example, Northern blot is often used to visualise differences in the quantity of mRNA produced by different groups of cells or at different times. In recent years, many new techniques have been developed to measure gene expression on a large-scale i.e. several thousands at once. Most of these techniques, including microarrays[18,19], RNA Sequencing (RNA-Seq)[20], reverse transcription polymerase chain reaction (RT-PCR)[21], serial analysis of gene expression (SAGE)[22], etc. work by measuring mRNA levels. The gene expression can also be analysed by directly measuring protein levels by Western blot[23], Protein Chips, Reverse Phase Protein Microarrays, Mass-Spectrometry based techniques[24], etc.

Among these techniques, DNA microarrays have been the most popular approach for transcript profiling until recently. However, array technology has limitations. For example, background hybridisation limits the accuracy of expression measurements, particularly for transcripts present in low abundance. In addition, arrays are limited to interrogating only those genes for which probes are designed. In recent

years, RNA-Seq technology, a high-throughput sequencing technology to directly sequence transcripts, is increasingly becoming a replacement to microarrays for whole-genome transcriptome profiling[25]. RNA-Seq has considerable advantages for examining a transcriptome's fine structure, such as the detection of unknown transcripts, allele-specific expression and splice junctions. It does not depend on genome annotation for prior probe selection and thus avoids the related biases present on microarrays. A recent study comparing between microarrays and RNA-Seq for transcriptome profiling showed a high correlation between gene expression profiles generated by the two platforms[26]. However, it also demonstrated that RNA-Seq is superior in detecting low abundance transcripts, differentiating biologically critical isoforms, and identifying genetic variants. In addition, RNA-Seq shows a broader dynamic range than microarrays, which allows for the detection of more differentially expressed genes with higher fold-change[26].

As transcriptome data used in this thesis were generated using microarray techniques, in the next sections only microarray techniques are described in detail.

1.3.2 Microarrays : cDNA and oligonucleotide

Microarrays are used to simultaneously detect relative expression of thousands of genes in the cellular pool of mRNA. DNA microarrays are microscope slides that are printed with thousands of tiny spots in defined positions, with each spot containing a known DNA sequence or gene. Often, these slides are referred to as gene chips or DNA chips. The DNA molecules attached to each slide act as probes to detect gene expession. There are different types of microarrays present based on the number and type of the probes, and on the number of channels (labelling colours) profiled on a single chip. The dual color complementary DNA (cDNA)-microarray was the prototype and its experimental process is represented in (Figure 1.3).

To perform a dual color DNA microarray analysis, mRNA molecules are typically collected from both an experimental sample and a reference sample. For example, the reference sample could be collected from normal cells, and the experimental sample could be collected from cancer cells. The two mRNA samples are then converted into cDNA, and each sample is labeled with a fluorescent probe of a different color. For instance, the experimental cDNA sample may be labeled with a Cy5(red) fluorescent dye, whereas the reference cDNA may be labeled with a Cy3(green) fluorescent dye. The two samples are then mixed together and allowed to bind to the microarray slide. The process in which the cDNA molecules bind to the DNA probes on the slide is called hybridisation. Following hybridisation, the microarray is scanned to measure the expression of each gene printed on the slide. If the expression of a particular gene is higher in the experimental sample than in the reference sample, then the corresponding spot on the microarray appears red. In contrast, if the expression in the experimental sample is lower than in the reference sample, then the spot appears green. Finally, if there is equal expression in the two samples, then the spot appears yellow. The data gathered through microarrays can be used to create gene expression profiles, which show simultaneous changes in the expression of many genes in response to a particular condition or treatment.

The most commonly used microarrays today are oligonucleotide arrays such as Affymetrix GeneChips[19]. These GeneChips are one-channel microarrays, meaning that the control and stress condition oligonucleotides are labeled with the same dye and each hybridised to a separate array (Figure 1.3). The probes

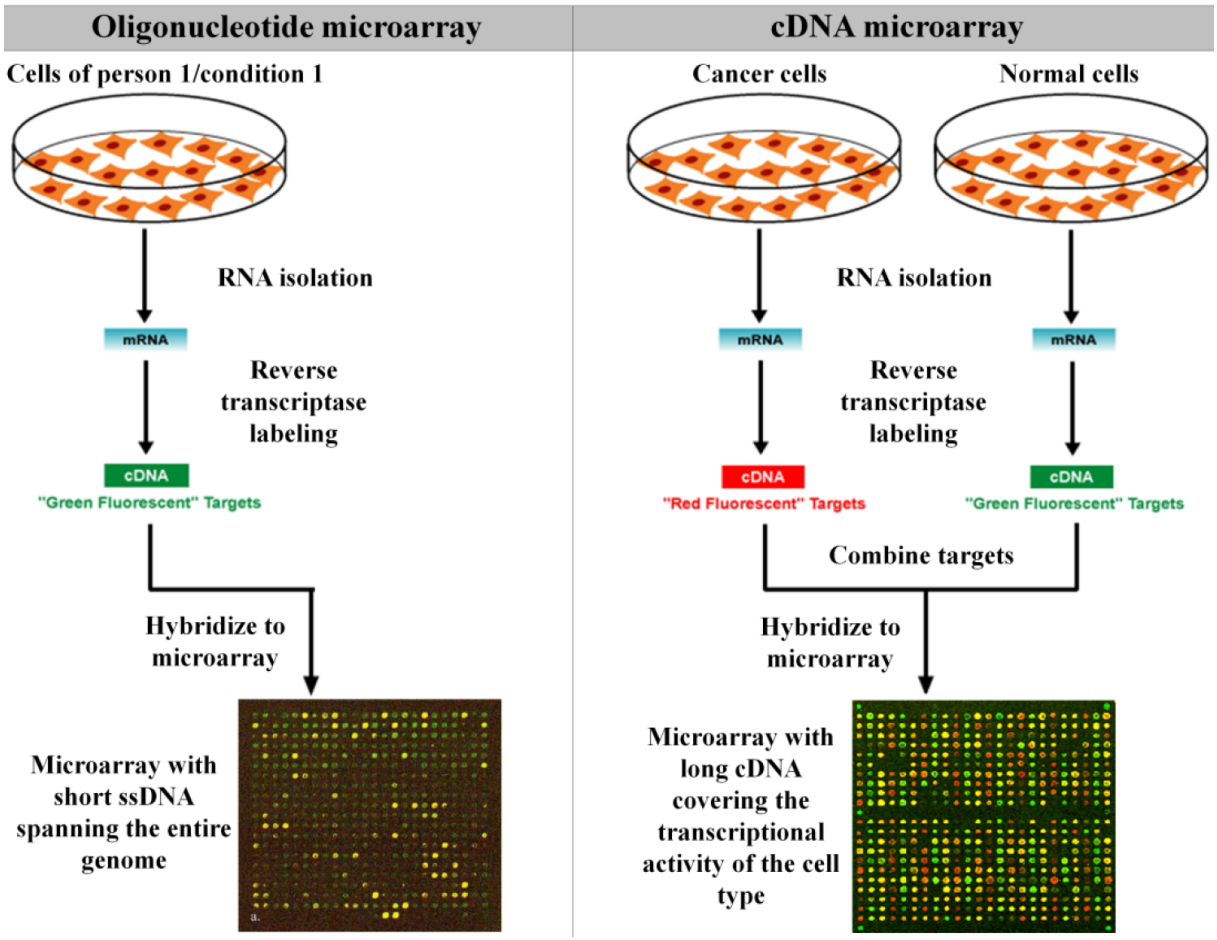


Figure 1.3: Schematised experimental process using a cDNA and oligonucleotide microarray. In dual-channel i.e. cDNA microarrays, RNA extracted from control and experimental sample are reverse transcribed into cDNA labelled with a fluorescent dye of a different color (here, green and red respectively). The two samples are then mixed together in equal amount and then hybridised on microarray slide. Finally, the microarray is scanned to measure the expression of each gene printed on the slide. In single-channel i.e. oligonucleotide microarrays, the control and stress condition oligonucleotides are labeled with the same dye and each hybridised to a separate array. Picture is taken from wikipedia commons.

on Affymetrix microarrays are short sequences of around 25 base pairs, and each gene is represented by a number of different probes attached to the microarray. The intensity of fluorescence is measured by a fluorescence scanner. In this PhD thesis, transcriptome data generated from ATH1[27] microarrays and AGRONOMICS1[28] tiling arrays (both single-channel) was used. The Affymetrix ATH1 microarray is frequently used for *Arabidopsis* transcriptomic analyses. The array consists of 22759 gene-specific probe sets, each containing eleven perfect match (PM) and eleven mis-matched (MM) probes (twenty-five base oligonucleotides hybridised to a glass slide). PM probes are complementary to the mRNA sequence; MM probes differ from the PM probes only at nucleotide thirteen, where the base is swapped to its complementary partner (e.g. C to G, A to T etc.). The array represents 22543 individual *Arabidopsis* loci (The *Arabidopsis* Information Resource release 8 (TAIR 8)), with some loci represented by more than one probe set.

As ATH1 (as well as other microarrays for transcript profiling) probe only about two-thirds of the annotated genes in the *Arabidopsis* reference genome, as an alternative a new tiling array, the AGRONOMICS1 Affymetrix tiling array[28] was developed to increase the genomic coverage. The probes on the AGRONOMICS1 array cover the entire nuclear *Arabidopsis* genome (TAIR 8), with the exception of repetitive sequences likely to cause cross-hybridisation. The AGRONOMICS1 array design doesn't

include the MM probes as it has been shown that the oligonucleotide microarray data can be robustly analysed based on PM probes only[29]. Additionally, the AGRONOMICS1 tiling array (in contrast to other available tiling arrays such as Affymetrix, Nimblegen, etc.) contains the complete paths of both genome strands, with on average one 25mer probe per 35-bp genome sequence window. Thus, this tiling array makes it possible to obtain strand-specific information on transcription units with a single array per sample and allows to investigate the correlation between chromatin states and expression. Overall, the AGRONOMICS1 tiling array yields very similar results to the ATH1 array in expression profiling experiments while providing advantages such as (i) information for many more genes (annotated as well as unknown genes), (ii) detection of alternative splicing and (iii) being compatible with ChIP-chip analysis.

1.4 Analysing and integrating transcriptome data

1.4.1 Preprocessing measurements

Gene expression studies are usually carried out for one or multiple samples (i.e. organs, tissues, cell types etc.) under different conditions and thus involve multiple microarrays. Before the measurements from these microarrays can be integrated into one analysis, the reported intensities (raw data) need to be preprocessed. The preprocessing generally involves (i) background correction to exclude false positive results arising due to local artefacts or unspecific probe binding, (ii) normalisation to allow comparison of different microarrays with each other and (iii) summarisation to integrate the intensities of different probes for a single gene. Although in recent years, many free and commercial packages have been developed to preprocess microarray data, the most commonly used algorithm is Robust Multi-array Average (RMA) algorithm[29–31]. In the background correction step, RMA assumes that the probe intensities are exponentially distributed and the background is normally distributed and greater than or equal to zero, to avoid negative values. The convolution background correction ensures that there are no negative values in the resulting corrected data. Further these values are transformed into $\log_2$ values to make them more human readable and make the normalisation additive. RMA offers a quantile normalisation algorithm in which, at first, the value of the highest intensity on each chip is replaced by the average of the highest intensities on each chip and next, the same is done for the second highest intensity and so forth. Eventually, the overall expression value distribution of each chip is comparable. Following quantile normalisation, an additive linear model is fit to the normalized data to obtain an expression measure for each probe on each microarray. The linear model for a particular probe can be written as

$$Y_{ij} = m_i + a_j + \epsilon_{ij} \tag{1.1}$$

where Y_{ij} represents the normalized probe value corresponding to the i^{th} GeneChip and the j^{th} probe within the probe set, m_i denotes the log-scale expression for the probe set in the sample hybridised to the i^{th} GeneChip, a_j denotes the probe affinity effect for the j^{th} probe within the probe set, and ϵ denotes a random error term. Tukey's median polish is applied to obtain estimates of the m_i values. For each row (microarray) the median is determined and subtracted from each value in that row. Afterwards the same is done for each column (probe set). This procedure is repeated until all row and column medians are

zero. The resulting values are subtracted from the original values. As such, column and row effects of the microarray are taken into account and because of the use of medians, outliers don't influence the resulting data. Finally the estimated m_i values serve as the log-scale expression measures associated with the particular probe set.

Once the data is preprocessed, the log-ratios of gene expression values in control and stress conditions can be calculated. These log-ratios are often used to identify the genes that significantly differ in expression between control and stress conditions, the genes with similar expression patterns, the enrichment of genes in a particular biological process, etc.

1.4.2 Differential gene expression analysis

One of the important tasks of analysing gene expression data is the need to identify genes whose expression patterns differ according to the experimental condition (stress vs control). A simple approach is to select genes based on fold-change criterion. This is usually only done when no replicates are available. However, this analysis doesn't allow the assessment of significance of expression differences in the presence of biological or experimental variation. Thus, when replicate measurements are available, statistical tests are used to test differential expression. Usually parametric tests such as *t*-test are used with the assumption that the underlying distribution is normal (or at least approximately normal). Since a large number of hypothesis tests are carried out (one for each gene), statistical analysis at a particular confidence level (e.g. 0.01 or 0.05) per gene may lead to a large number of false positive results. Two multiple testing corrections are commonly used to control the number of false positives: the Bonferroni correction[32] which controls the Family Wise Error Rate (FWER) and the Benjamini and Hochberg correction[33] which controls the false discovery rate (FDR). The FWER is the probability of making at least one Type I error (i.e. one false positive, rejecting the null hypothesis though it is true). The FDR correction controls the rate of falsely rejected null hypotheses when being true (false positive rate).

In this thesis, limma[34] (linear models for microarray data, a bioconductor[35] package) was used to identify differentially expressed genes. Limma first fits a linear model to analyse complex experiments with multiple treatment factors and uses quantitative weights to account for the variation in precision between different observations. Later, limma uses moderated *t*-test statistics to determine whether gene expression values in the stress and control conditions significantly differ from each other, under the assumption that the log-ratios are normally distributed. The resulting *t*-test statistic is used to compute a P value for the observed expression ratio, under the null hypothesis that the mean expression levels of the gene under stress and control conditions are the same. The null hypothesis is rejected for P values below a given significance level (for e.g. 0.01 or 0.05). Limma further corrects P values for multiple testing, for instance using the Benjamini-Hochberg method[33] at a specified FDR threshold (e.g. 0.01, 0.05, etc.). The typical outputs from limma analysis are a log-ratio matrix and an associated FDR corrected P value matrix. Besides, the P value matrix is often converted into a differential expression matrix, which is generally used as an input for (bi)clustering algorithms. The matrix usually contains values 1 or 0 representing respectively differentially and non-differentially regulated gene expression in the stress condition compared to control. In this context, for a non-significant P value i.e. no differential expression

between stress and control, the value 0 is given in a differential expression matrix. Otherwise, the value 1 is given in a differential expression matrix.

1.4.3 Gene ontology (GO) enrichment analysis

The differentially expressed genes are usually studied for their enrichment in a particular biological process. These kinds of analyses are greatly facilitated by a structural description of known biological information at different levels of granularity. The GO project was initiated in 1996[36] with the aim to capture the increasing knowledge of gene function in a controlled vocabulary applicable to all organisms. GO consists of three hierarchically structured vocabularies that describe gene products in terms of their associated biological processes, molecular functions and cellular components. GO has the structure of a directed acyclic graph; each node representing a GO term, each edge representing a linking phrase: 'is a', 'part of' or 'regulates'. 'Regulates' is divided in 'positively regulates' and 'negatively regulates'. Every term (node) is given a GO identification number that makes it easy to replicate the search.

In the past years, many tools have been developed that analyse GO term enrichment in a given gene set. A comprehensive list can be found at (`http://www.geneontology.org/GO.tools.shtml`). The gene-to-annotation format of GO allows these tools to systematically map genes in a given list to the associated GO terms and then statistically examine the enrichment of gene members for each of the GO terms by comparing the outcome to reference background. A recent survey by Huang and co-workers[37] classified some of these tools into three classes: (i) singular enrichment analysis (SEA), (ii) gene set enrichment analysis (GSEA) and (iii) modular enrichment analysis (MEA), based on the algorithms used. SEA based tools (GoStat[38], GoMiner[39], BiNGO[40], DAVID[41], etc.) use the most traditional strategy for enrichment analysis that takes the user's preselected genes and then iteratively test the enrichment of each GO term one-by-one in a linear mode. Then, enriched terms are listed in a simple linear text format. The enrichment P value calculation is performed by common and well-known statistical methods, including Chi-square, Fisher's exact test, Binomial probability, etc and corrected for multiple testing. This class is capable of analysing any gene list that could be selected from any high-throughput biological studies. However, the limitation of this class of algorithms (except for tools such as BiNGO that additionally output GO Hierarchy coloured based on the significance values) is that the deeper inter-relationships among the terms may not be fully captured in a linear format report. GSEA based tools (GSEA[42], PAGE[43], GO-Mapper[44], etc.) consider all genes (without any pre-selection) and use associated experimental values (fold change) in the enrichment analysis. The unique benefit of this strategy is that no prior arbitrary cutoff's are needed for gene selection and additional experimental values are integrated into P value calculation. The output of this strategy is the maximum enrichment score (MES), which is calculated from the rank order of all gene members in the GO term. The enrichment P value calculation is performed by methods that includes randomisation approaches, Kolmogorov–Smirnov-like statistics or parametric statistical approaches such as z-scores, t-tests, etc. This class is suitable for pair-wise biological studies, for instance perturbation vs. control. The limitation of this class of algorithms is that it may be difficult to apply to the diverse data structures derived by complex experimental design and new technologies. MEA based tools (ADGO[45], Ontologizer[46], GoToolBox[47], etc.) use the basic enrichment algorithms found in SEA and incorporate extra network discovery algorithms by considering the gene-to-gene (or

term-to-term) relationships. The enrichment P value is calculated by measuring enrichment on joint terms or by considering parent-child relationships or by measuring term-term global similarity with Kappa Statistics, Pearson's correlation, etc. This class is capable of analysing any gene lists like SEA class but the interesting orphan genes/terms (with little relationships to other genes/terms) may be left out from the analysis.

In this thesis, BiNGO[40] (Biological Networks Gene Ontology tool) was used to assess the over-representation of a GO categories in a set of genes. A major advantage of BiNGO over other ontology enrichment tools is the flexibility of using custom ontologies and annotations. Next to the standard Gene Ontology, it is possible to use GOslim ontologies or a user-defined ontology file. The BiNGO tool also offers a choice of statistical tests (namely hypergeometric test, with an exact P value as result or the binomial test, resulting in an approximate P value) for the analysis. The hypergeometric test calculates the probability of a random set of genes being enriched in a particular function when sampling without replacement. For each GO category, a P value is given that represents the probability whether or not the observed number of genes annotated to that GO term in the gene set under study is generated by chance. The binomial distribution probes the same probability as the hypergeometric distribution but through sampling with replacement. The BiNGO tool further corrects the P values for multiple testing (FDR or FWER), as one analysis requires the testing of all gene ontology terms for significant enrichment in that gene set.

1.4.4 Gene co-expression network analysis

Differential expression relationship among genes across various conditions are greatly studied to elucidate the functional relationship among genes. In this context, gene expression profiles of a number of experimental conditions (combined together in a compendium) are categorised into clusters based on the similarity in their patterns of expression. The functions of unknown gene products in a cluster are then inferred using deduction techniques based on the so-called guilt-by-association principle[48]. The co-expressed genes in each cluster are inferred to be coding for proteins that participate in a common biological function. There are two classic techniques used on gene expression data for predicting or annotating gene functions. The first technique is a form of unsupervised learning called 'clustering', while the second is a form of supervised learning called 'classification'[49].

Clustering methods assume to have no prior knowledge about any of the genes' biological functions and use the expectation that genes that perform a common biological function would have expression profiles that exhibit a similar pattern across different experimental conditions. The clustering process organises genes into different functional classes using a similarity (or distance) measure (e.g. Pearson correlation, Euclidean distance, mutual information, etc.) on the gene expression data. In recent years, many clustering techniques[50] have been developed to find clusters of co-expressed genes. These techniques include hierarchical clustering[51], k-means clustering[52], simulated annealing-based clustering[53], graph-theoretic clustering[54], etc. In the classification method, prior knowledge (e.g. subset of genes involved in a biological pathway of interest) is exploited in the form of training sets for supervised machine learning algorithms to identify unknown genes belonging to the similar functional classes. Several classification methods have been developed, including pattern discovery methods[55], support vector machines[56]

and neural networks[57]. In addition, combined approaches of these two classes of techniques are present for assigning biological functions to the unknown genes[58].

The traditional clustering methods (mentioned above) are well suited for time series expression data to capture global tendencies of co- or anti-regulation between genes. However, they are not appropriate for perturbation experiments, as genes are not necessarily co-expressed under all experimental circumstances or they may be co-regulated under some perturbations and show uncorrelated or even inversely correlated expression behaviour under some experiments[1]. Thus, an alternative clustering strategy referred to as biclustering has been developed for specifically detecting subsets of genes that exhibit similar behaviour across a subset of conditions. In recent years, a number of biclustering methods have been developed[59,60], which can be broadly categorised based on the type of biclusters investigated and the underlying mathematical formulation used to investigate them[61]. Very recently, Oghabian and coworkers (2014)[61] reviewed and further categorised biclustering techniques into four classes such as (i) Correlation maximisation biclustering methods (CMB), (ii) Variance minimisation biclustering methods (VMB), (iii) Two-way clustering methods (TWC) and (iv) Probabilistic and generative methods (PGM). CMB based tools (such as ACV[62], BiMine[63], CC[64], FLOC[65], etc.) seek for subsets of genes and samples where the expression values of the genes or respectively samples correlate highly among the samples or respectively genes. For instance, CC (Chen and Church)[64] used mean squared residue score threshold to find biclusters, where this score was used as a measure of the coherence of the genes and conditions in the bicluster. VMB based tools (such as BiMax[66], Spectral[67], XMOTIF[68], etc.) search for biclusters in which the expression values have low variance throughout the selected genes, conditions or the whole submatrix (of gene by conditions). For example, XMOTIF[68] extracts biclusters (or conserved expression motifs) with genes exhibiting constant expression for a subset of samples. TWC based tools (such as CTWC[69], ISA[70], ITWC[71], etc.) discover homogeneous biclusters by iteratively performing one-way clustering on the genes and samples. For example, coupled two-way clustering (CTWC)[69] approach iteratively search subsets of the genes that remain constant through the iterations of the algorithm and are used as the attributes for the clustering of the samples and *vice versa*. PGM based tools (such as CMonkey[72], FABIA and FABIAS[73], Gibbs biclustering[74], Plaid[75], SAMBA[76], etc.) employ probabilistic techniques to discover genes or samples that are similarly expressed across a subset of samples or genes respectively in the given data-matrix. For example, SAMBA (Statistical-Algorithm Method for Bicluster Analysis)[76] is a graph theoretic algorithm coupled with statistical modelling of the data, in which the input expression data is modelled as a bipartite graph whose two parts corresponds to conditions and genes respectively. The algorithm basically looks for an optimal set of heavy-weighted subgraphs (biclusters) that covers this bipartite graph. Each of the methods described above have their strengths and weaknesses. Some of the methods such as CC[64] and Gibbs sampling[77] are less suited to find overlap between biclusters as they mask previously found biclusters with random noise or because methods like e.g. Spectral[67], CTWC[69], etc. partition the data. Other methods require user input about the desired number of biclusters in advance or generate highly redundant biclusters, require extensive adjustment of parameters or do not integrate other types of biological data such as differential expression analysis information (P values) from perturbational data, etc.

In this thesis, ENIGMA (Expression Network Inference and Global Module Analysis)[1], a graph-based biclustering like method, was used that tackles some of these issues. The ENIGMA algorithm

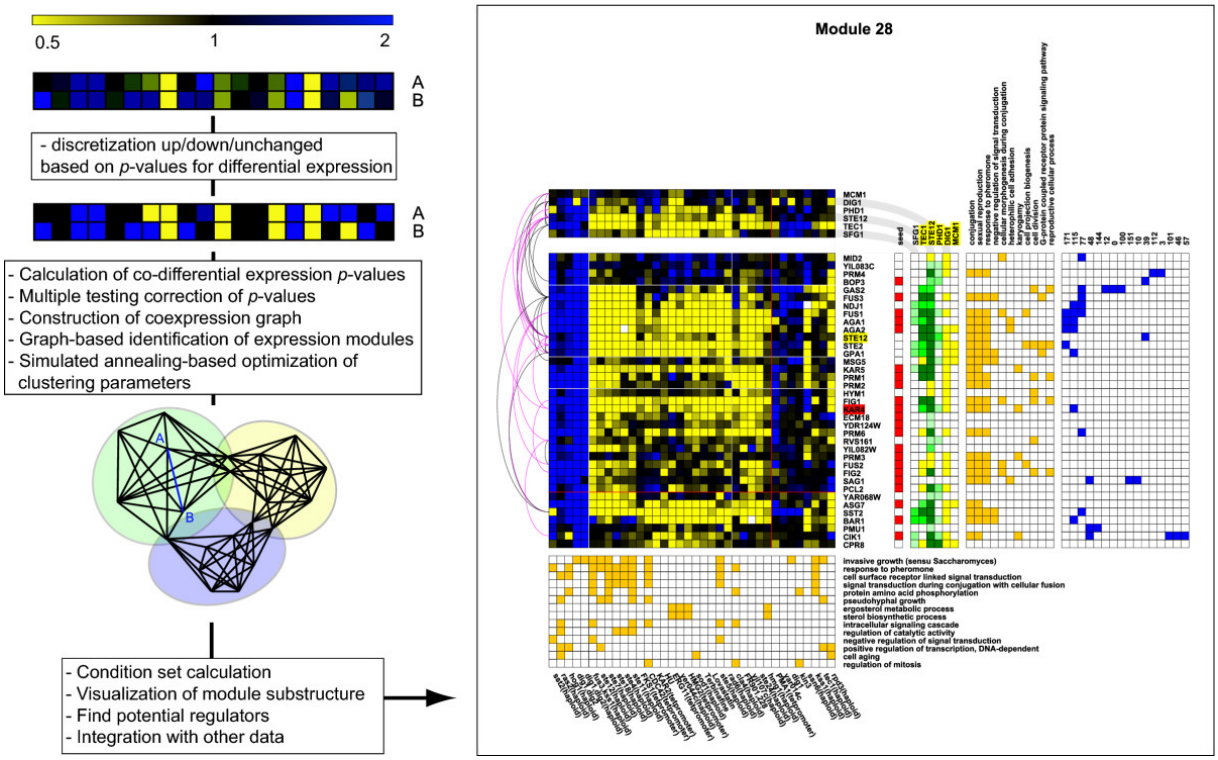


Figure 1.4: An overview of ENIGMA algorithm. The input gene expression values are discretised into three categories i.e. up-regulated, down-regulated and unchanged based on their differential expression P value. If the gene is significantly up-regulated in a given experiment, the corresponding field is labeled blue. Experiments in which the gene is significantly down-regulated are similarly labeled yellow, and the remaining fields are labeled black. The exact overlap of yellow and blue field positions between profiles (A and B) is then tested for significance with the null hypothesis that the response of the genes A and B to the perturbations are uncorrelated. The resulting significant co-differential expression P values are corrected for multiple testing and translated to edges in a co-expression network, which is clustered into expression modules (i.e. groups of significantly co-differentially expressed genes) using a graph-based clustering algorithm. The module (Right panel) is split in leafs in both dimensions based on average linkage clustering using a cosθ threshold of 0.65. Red lines (rightmost leaf and bottom leaf) indicate leafs (Beyond the red line) of size < 3 grouped in a single leaf. Transcription factors are highlighted in yellow in the gene list if there is ChIP data available for them, while other regulators are highlighted in red. To the right of the expression matrix is a column indicating the module's seed genes (red). Further to the right is a matrix depicting the presence of enriched TF binding sites (yellow) and/or significant co- or anti-expression links with potential regulators (green and red, respectively; the hue is proportional to the P value of the link; in case of overlap with an enriched binding site, the field is coloured dark green or dark red). The expression profiles of these regulators are depicted on top of the module's expression matrix. To the far right are matrices depicting the genes' membership of enriched GO categories (orange) and membership of other modules (blue). The black and magenta arcs represent protein and genetic interactions, respectively. This figure is reproduced from Maere et. al.[1]

(Figure 1.4) uses (as input) the expression log-ratios and differential expression P values [i.e. 0(no-differential regulation), 1(up-regulation), -1(down-regulation)] from any perturbational (chemical, genetic or environmental) microarray dataset as input, and extracts gene expression modules (as an output) based on the use of combinatorial statistics and graph-based clustering. The ENIGMA algorithm is able to detect significant partial co-differential expresssion relationships between genes and overlap between modules. ENIGMA further characterises the obtained modules by incorporating other data types, e.g. GO annotation, protein interactions and transcription factor binding information, and suggests regulators (i.e. genes that are significantly more connected to a module, through positive or negative co-expression edges, than expected at random tested using hypergeometric test, default FDR = 0.05, selected either from a user-defined list or a user-defined set of GO classes) that might have an effect on the expression of (some of) the genes in the module. The combinatorial statistic used by ENIGMA assesses which pairs of genes are significantly co-expressed (co-differentially expressed) across at least a subset of conditions. The resulting co-expression P values are then corrected for multiple testing and translated to edges in a

co-expression network, which is further clustered into expression modules (i.e. groups of significantly co-differentially expressed genes) using a graph-based clustering algorithm[1].

1.5 Modeling approaches for gene function inference

Gene function determination is a critical step towards understanding the biological processes of a system and the system itself. Although, the traditional laboratory methods used to determine gene functions, such as forward or reverse genetics, are accurate and reliable, they require enormous effort and time. Thus, with the increase in sequenced genomes and availability of transcriptome profiles, many computational approaches have been developed to predict gene functions in a timely and cost-effective manner, with the aim of further guiding laboratory experiments and facilitating more rapid functional annotation of genomes.

Among these predictive approaches, the most widely used are homology-based methods that take into account similarities in DNA or protein sequences for gene function inference. The rationale for the homology-based method is that two proteins with a similar sequence or structure have evolved from a common ancestor and thus may have similar functions. However, it is likely that homologous proteins might also acquire different functions in evolution. Hence, in recent years, researchers have developed various methods that use microarray expression data[78], protein domain configuration[79], protein-protein interaction networks[80], and phylogenetic profiles[81] to predict functions of genes. Gene function inference from different types of biological data simultaneously integrated into functional networks and modules, are attracting more attention as these methods have been shown to yield more accurate predictions[82–84]. In this context, functional networks are usually constructed with nodes corresponding to genes and edges representing the co-functionalities of gene pairs and used as input for learning algorithms. There are two main approaches that use functional networks to predict gene functions: (i) network-based (direct annotation), which infer the function of genes based on its connections in the network, and (ii) module-based, which first identify modules of related genes in the network (or directly from the data) and then annotate each module based on the known functions of its members[85].

The module-based approach is described in the section 1.4.4. Here, two network-based approaches for the plant model organism *Arabidopsis* are explained in brief. The first approach allows to predict functions of genes at a system level, whereas the second allows to predict the gene functions at the level of a tissue or developmental stage of interest. (i) AraNet is a probabilistic functional gene network, which was constructed to attribute novel functions to *Arabidopsis* genes[86]. This network was constructed by integrating diverse 'omics' data such as mRNA co-expression patterns, protein-protein interaction data, genomic contexts of orthologous proteins, protein domain co-occurrence profiles and functional linkage data transferred from other organisms (e.g., *C. elegans*, *Drosophila melanogaster*, *Homo sapiens* and *Saccharomyces cerevisiae*) by orthology relationships. Each interaction in the network has an associated log-likelihood score that measures the probability of an interaction representing a true functional linkage between two genes. This network was used by the authors[86] to predict candidate genes associated with the set of 23 known embryo pigmentation genes. Out of the suggested 200 genes, 90 genes were tested using homozygous transfer DNA (T-DNA) insertional mutant lines. A total of 14 genes exhibited color and

morphology defects in young seedlings, reminiscent of embryo pigmentation mutants. This study overall represented a tenfold enrichment in the discovery rate of the mutant phenotype over that observed during a forward-genetics screen of T-DNA insertion lines[86]. (ii) Many *Arabidopsis* gene products are known to be functional in a specific tissue or during a specific developmental period[2]. Thus, investigating gene functions experimentally on every plant structure at each of its development stages individually would be tedious and costly. Therefore, a compendium of probabilistic functional networks was constructed by integrating over 60 microarray, physical and genetic interaction, and literature curation datasets[87]. This compendium includes tissue-, biological process-, and development stage-specific networks, which were inferred using Bayesian classifiers for heterogenous data, each predicting relationships specific to an individual biological context. These functional networks for e.g. disease resistance, root hair patterning, and auxin homeostasis have been shown to yield for reliable gene function predictions. Moreover, they enable the rapid investigation of uncharacterised genes in specific tissues and developmental stages of interest.

In this thesis, we used PiNGO[88], a network-based method, to find genes associated with processes or pathways of interest. The PiNGO tool allows the user to load a network, e.g. a gene co-expression, protein or genetic interaction, or integrated network either through Cytoscape[89] or from a text file. The PiNGO algorithm screens a particular network for genes whose direct neighbours are enriched for given GO categories (i.e. "start" GO categories) along with its subcategories at a chosen significance level (i.e. 0.01 or 0.05). Simultaneously, PiNGO can exclude genes with certain functional properties from the analysis (i.e. "filter" GO categories), or focus on genes with particular functions (i.e. "target" GO categories). Similar to the BiNGO tool, PiNGO uses hypergeometric or binomial tests to calculate enrichment statistics, and Bonferroni or Benjamini–Hochberg FDR corrections to adjust the resulting P values for multiple testing. The gene co-expression networks underlying gene function prediction approaches are generally constructed from traditional gene profiling experiments that involve relatively harsh perturbations on pooled samples. In the next Chapter, we discuss an alternative approach, which involves a network constructed from data sets of multiple subtle perturbations to the systems on individual plants.

1.6 Author contributions

I wrote this chapter by myself. It resulted from the many fruitful discussions with both my promoters.

Chapter 2

Predicting gene function from uncontrolled expression variation among individual wild-type *Arabidopsis* plants

Rahul Bhosale*, Jeremy B. Jewell*, Jens Hollunder, Abraham J.K. Koo, Marnik Vuylsteke, Tom Michoel, Pierre Hilson, Alain Goossens, Gregg A. Howe, John Browse and Steven Maere. *Plant Cell* **25**:2865-2877.
*These authors contributed equally to this work.

Abstract

Gene expression profiling studies are usually carried out on pooled samples grown under tightly controlled experimental conditions, to suppress variability among individuals and increase experimental reproducibility. In addition, to mask unwanted residual effects, the samples are often subjected to relatively harsh treatments [i.e. treatment that moves cells far from its normal working point; treatments such as chemical (reagents, mutagens, etc.), genetic (gene deletion, knockdown or over-expression) or environmental (desiccation, high or low temperatures, etc.)] that are unrealistic in a natural context. Here, we show that expression variations among individual wild-type *Arabidopsis* plants grown under the same macroscopic growth conditions contain as much information on the underlying gene network structure as expression profiles of pooled plant samples under controlled experimental perturbations. We advocate the use of subtle uncontrolled variations in gene expression between individuals to uncover functional links between genes and unravel regulatory influences. As a case study, we use this approach to identify *ILL6* gene as a new regulatory component of the jasmonate response pathway.

For the author contributions, see page 47.

2.1 Introduction

A classical dogma in systems biology states that in order to study a biological system, one needs to systematically perturb the system, measure the response and construct a model that predicts the outcome of future perturbations[90]. For instance, molecular biologists often profile the mRNA expression response to controlled perturbations, such as environmental or chemical treatments or genetic knockouts. Because reproducibility is a cornerstone of the scientific method, such experiments are invariably performed in a tightly controlled setup[91]. Great care is taken to control the boundary conditions and to keep unwanted external influences in check. Variability among individuals is smoothed out by pooling biological materials and averaging over biological replicates. Moreover, in order to overpower any residual uncontrolled effects, the perturbations applied to the system under study are often rather harsh, causing the system to operate outside its normal range.

Even when taking such precautions, the reproducibility of expression profiling experiments is often poor, in part because reproducing particular experimental conditions is hard even when detailed information on the original setup is available[3]. To assess the within- and between-lab reproducibility of leaf growth-related (molecular) phenotypes, Massonnet and co-workers[4] recorded the gene expression profiles of 41 individual leaves at the same developmental stage (leaf 5, stage 6.0), taken from *Arabidopsis thaliana* plants of three accessions (Col-4, Ler, Ws) grown in six different laboratories. Despite the fact that the participating labs adhered to a standardised and very detailed protocol, significant intra- and inter-laboratory variability in gene expression was found. The authors concluded that small variations in growth conditions within and across labs may lead to substantially different gene expression profiles.

The key question addressed in this study is whether we can use such uncontrolled expression variations to our advantage in a reverse engineering context, i.e., to unravel the wiring of an organism. We reanalyse the gene expression data set of Massonnet and co-workers[4] and compare its functional prediction performance to that of same-sized compendia of *Arabidopsis* gene expression experiments profiling the response to controlled perturbations on pooled plant samples. We show that, from a guilt-by-association perspective, subtle uncontrolled variations among individual leaves are as informative as experiments monitoring more severe controlled perturbations in pooled samples. Since it is often practically infeasible to define and perform the tens to hundreds of controlled perturbations needed to unravel (part of) a transcriptional regulatory network, our findings may open up novel avenues to generate sufficient amounts of data for reverse engineering algorithms.

2.2 Results

2.2.1 Residual gene expression differences yield biologically relevant expression modules

The gene expression data set of Massonnet and co-workers[4] contains expression profiles of leaves of three accessions grown in six different labs (Table 2.1), which causes a substantial proportion of the expression variance among leaves to result from lab and accession effects (Figure 2.1). Accession, lab and lab $\times$ accession effects explain on average 14.9, 19.7 and 12.8% of the expression variance of a

single gene, respectively, whereas the residual error contains 52.5% of the variance on average (median values 9.9, 17.0, 11.4 and 53.8%, respectively). Although the variance induced by lab or accession effects may well contain biologically relevant information, we were primarily interested in analysing the gene expression variation among comparable individual plant leaves grown under comparable macroscopic growth conditions. Substantial lab and accession effects, by virtue of not being independent and highly redundant across the leaves profiled, are expected to largely overpower the residual variation of interest when calculating co-expression links (see below). Therefore, we used a two-way unbalanced design analysis of variance (ANOVA) model to remove lab, accession and lab $\times$ accession effects from the data set (Section 2.4). The residuals of this ANOVA analysis (i.e. the unexplained expression differences among the 41 individual leaves, further referred to as the residuals data set) are the basis of all following analyses.

Table 2.1: Number of individual leaves profiled per lab and accession in the Massonnet et al. (2010) study. The codes P1, P2A, P2B, P3A, P3B and P4 denote the six labs that contributed the leaf samples that were expression profiled. The identity of the labs corresponding to particular codes was not disclosed in the original study[4]. All leaves were expression profiled in the same lab. The table indicates the number of samples on which unbalanced design ANOVA estimation of lab- and accession-effects was based in the rightmost column and bottom row, respectively. Estimation of lab x accession effects was based on the numbers of samples indicated in the core table.

	col-4	*Ler*	*Ws*	Total
P1	3	-	-	3
P2A	3	3	3	9
P2B	3	3	3	9
P3A	3	-	3	6
P3B	3	-	3	6
P4	3	3	2	8
Total	18	9	14	41

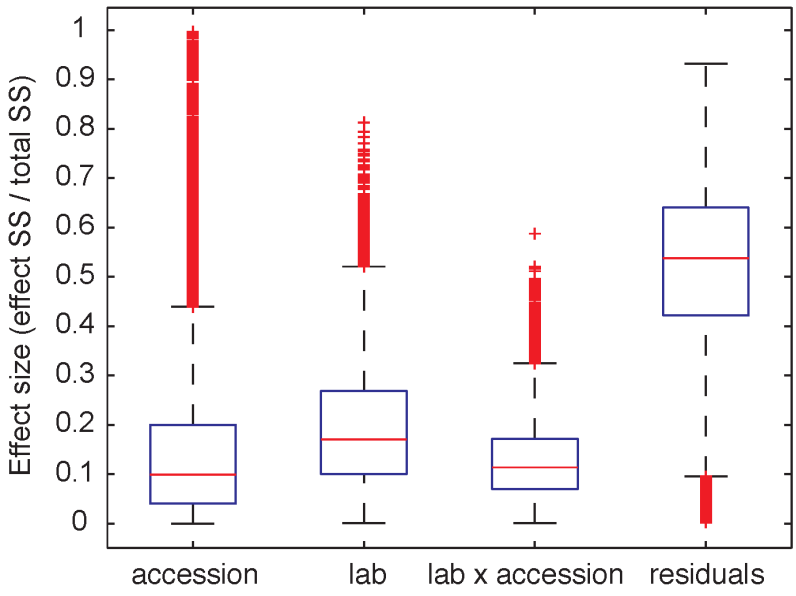


Figure 2.1: Effect size of accession-, lab-, lab x accession- and residual effects in the Massonnet et al. (2010) data set. Gene expression effect sizes for 19,760 genes were calculated as η = Effect Sum of Squares (SS)/ Total Sum of Squares (SS), with SS estimated through two-way unbalanced design ANOVA analysis in MATLAB®(`http://www.mathworks.com/`, anovan function with 'sstype'=1) and GenStat®(`http://www.vsni.co.uk/software/genstat`), yielding identical results. Boxes extend from the 25th to the 75th percentile, with the median indicated by the red line. Whiskers extend from each end of the box to the most extreme values within 1.5 times the interquartile range from the respective end. Data points beyond this range are displayed with a red + sign.

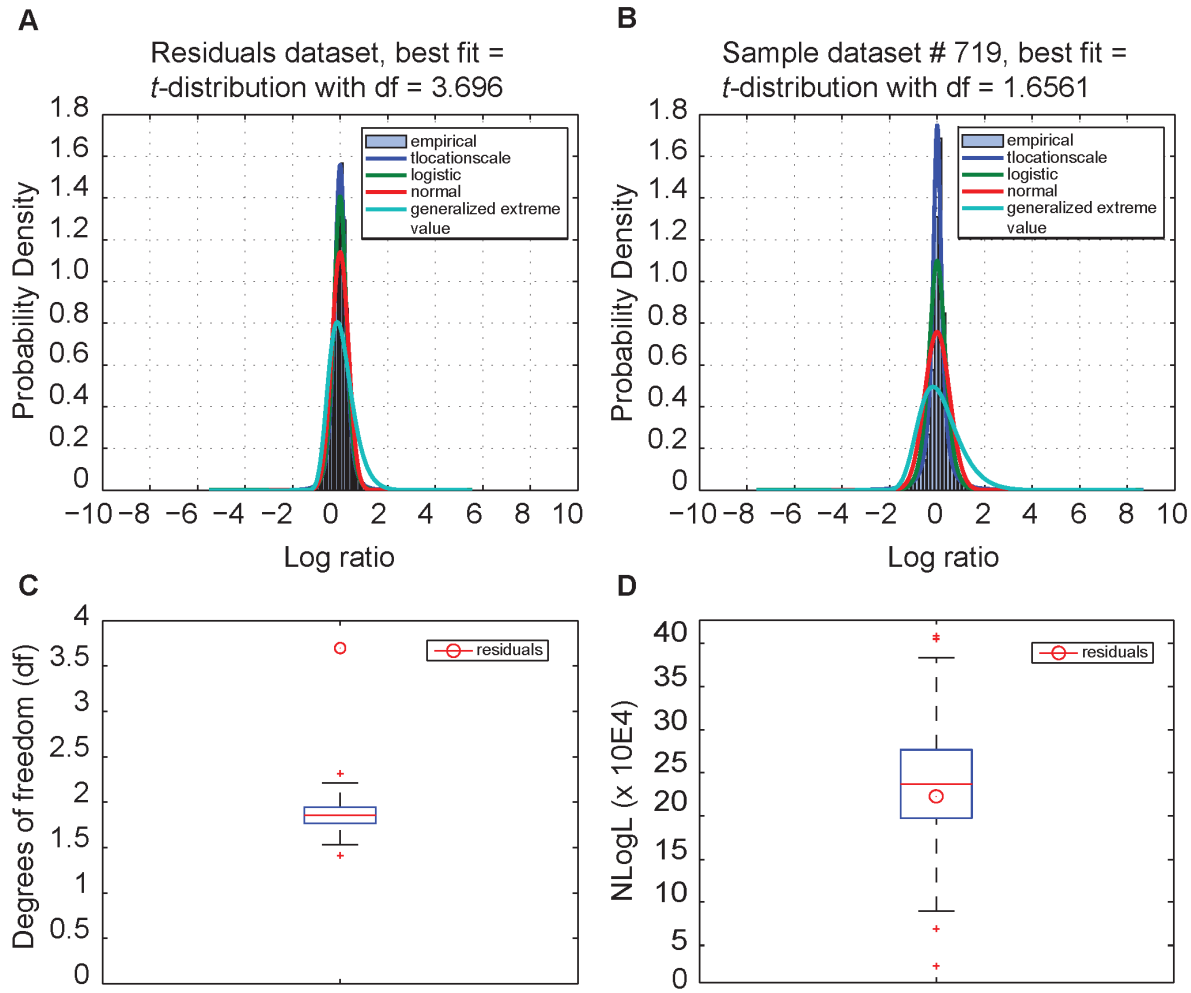


Figure 2.2: Distributional characteristics of log-ratio expression values in the residuals and sample data sets. (**A.**) Log-ratio expression value distribution for the residuals data set, best fit by a location-scale *t* distribution with df = 3.696. (**B.**) Log-ratio expression value distribution for a representative sample data set, best fit by a location-scale *t* distribution with df = 1.6561. (**C.**) Box-and-whisker plot of the best-fit *t* distribution degrees of freedom (df) parameter for all 1000 sample data sets. The residuals distribution has a substantially higher df parameter than the sample data sets, indicating that it is somewhat closer to a normal distribution (df = ∞). (**D.**) Box-and-whisker plot of the negative log likelihood of the *t* distribution fits to the observed data distributions. Boxes extend from the 25th to the 75th percentile, with the median indicated by the red line. Whiskers extend from each end of the box to the most extreme values within 1.5 times the interquartile range from the respective end. Data points beyond this range are displayed with a red + sign.

We used the ENIGMA algorithm[1] to calculate expression modules from the residuals data set and 1000 randomly assembled compendia of 41 gene expression profiles of controlled perturbational treatments on pooled *Arabidopsis* leaf or shoot material (referred to as the sample data sets; see Section 2.4). The log-scaled residuals data set is best fit by a Student's *t* location-scale distribution with a *df* parameter of 3.70, whereas the sample data sets exhibit a *t* distribution with *df* in the range 1.41 to 2.31, indicating that the log ratio distributions of the sample data sets contain somewhat heavier tails (i.e. more expression values that are substantially up- or down-regulated with respect to the normal expectation) (Figure 2.2). This may not come as a surprise given that the sample data sets include experiments profiling gene expression responses to major-effect perturbations, as opposed to the residuals data set. The ENIGMA algorithm requires discretisation of expression values into the categories "up-regulated", "down-regulated", and "unchanged" (or "undecided")[1]. The algorithm was originally intended for detecting significant "co-differential expression", a hybrid measure between co-expression and differential expression that essentially indicates whether two genes are significantly up-

or down-regulated together over at least a subset of the conditions profiled. The underlying rationale is that simple co-expression measures, such as Pearson's correlation, may be misleading in cases where co-regulated genes respond qualitatively the same, but quantitatively different to a series of different regulatory inputs. Discretisation of the gene expression response into up/down/unchanged removes some of the quantitative disturbances that may obfuscate co-expression patterns and allows for the use of combinatorial statistics to assess significant co-differential expression relationships over part of the condition set instead of the entire set[1]. Since statistically motivated differential expression P values can only be computed for perturbational data sets with biological replicates, such as the sample data sets, but by design not for the residuals data set, we used a uniform log ratio threshold instead to define up- and down- regulated gene expression values in all data sets. Therefore, "differential" expression in this context is not motivated in terms of statistically rigorous differential expression P values, but merely serves as a means to discretize the expression values for ENIGMA analysis and to separate noise (technical noise and some forms of intrinsic stochastic noise) from potentially valuable signal. All mentioning of "differential" expression in the remainder of the article should be interpreted accordingly. For thresholds in the appropriate range (i.e. before the distribution tails start flattening out), the residuals and sample data sets contain numbers of differential log ratio expression values in the same range. We fixed the $\log_2$ ratio threshold at 0.3498 (i.e. the standard deviation of the residuals data set), corresponding to a fold change threshold of 1.274 (Figure 2.3).

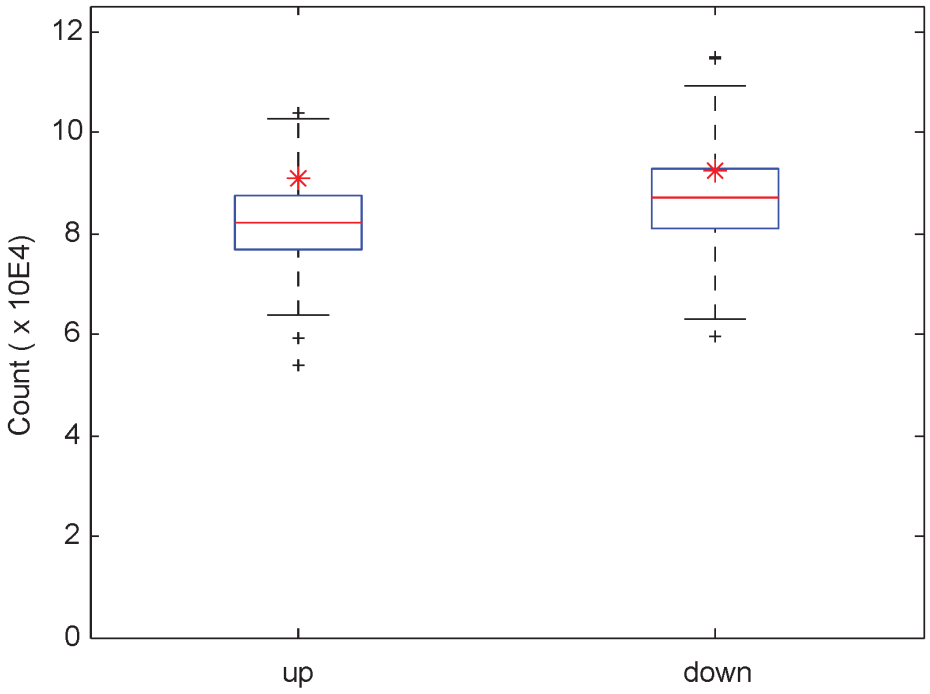


Figure 2.3: Numbers of 'differential' expression values in the residuals and sample data sets, for the purpose of ENIGMA analysis. 'Differential' expression was assessed using a uniform $\log_2$ ratio cutoff of 0.3498 for all data sets (equal to one standard deviation for the residuals data set), corresponding to 1.274-fold up- or down-regulation. Note that 'differential' expression in the present context is not motivated in terms of statistically rigorous differential expression P values (which cannot be computed for the residuals data set), but that the log ratio cutoff merely serves as a means to discretise the expression values for ENIGMA analysis. A cutoff value in the appropriate range ensures that the sample and residuals data set have comparable numbers of up- and down-regulated expression values. Box-and-whisker plots summarise the count distributions over the 1000 sample data sets, and the red stars indicate the counts for the residuals data set. Boxes extend from the 25th to the 75th percentile, with the median indicated by the red line. Whiskers extend from each end of the box to the most extreme values within 1.5 times the interquartile range from the respective end. Data points beyond this range are displayed with a black + sign.

Interestingly, we observed that the residuals data set still provides enough signal to discriminate biologically relevant expression modules (Figure 2.4; Supplemental Data Sets 1 and 2). Sister plants (same lab, same ecotype) often exhibit different residual expression responses in a given module (i.e. the module genes are up-regulated in one sibling plant and down-regulated in another; Figure 2.4), indicating that the modules are not formed by lingering lab or accession effects that were not removed by ANOVA analysis (i.e. effects that are nonlinear on log scale; see Section 2.4), in contrast with many of the modules learned from the original data set (Supplemental Data Sets 3 and 4). The set of modules learned from the residuals data set contains modules that are significantly enriched in, among other processes, photosynthesis, ribosome and chromatin assembly, proteolysis, secondary metabolism, response to wounding, bacteria, chitin and jasmonic acid (JA) stimulus, response to temperature, water, and nutrient levels, and starch catabolism (see Section 2.4 and Supplemental Data Set 2). The fact that the recovered modules are enriched for a variety of biological processes indicates that the residuals are not merely noise, but are to a large extent defined by genuine differences in the expression response of particular regulons, presumably caused by subtle uncontrolled variations in the growth conditions of individual plants (see below).

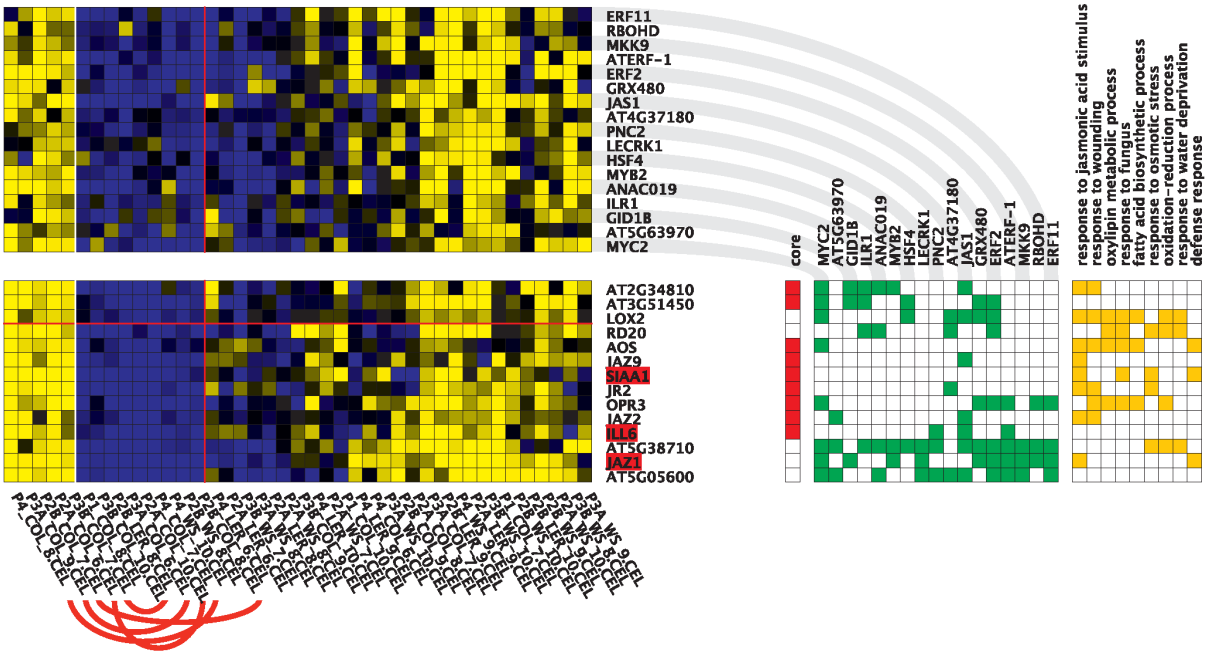


Figure 2.4: Co-differential expression module enriched for 'response to JA stimulus' genes, obtained with ENIGMA [1] on the residuals data set. Yellow/blue squares indicate up-/down-regulated gene expression with respect to the baseline leaf expression of the gene concerned. The bottom matrix contains the expression profiles of the module genes, while the top matrix contains the expression profiles of predicted regulators of the module. Significant co-differential expression links between the regulators and the module genes are indicated in the green matrix to the right. Genes highlighted in red are regulators that are part of the module. Genes indicated as core genes belong to the original module seed, other genes were accreted by the seed in the course of module formation [1]. Gene annotations for enriched GO categories are indicated in the orange matrix to the right. Sister plants (same lab, same ecotype, indicated by red arcs for the first two condition leaves) often end up in different condition leaves in the module, indicating that expression variations between individual plants, and not residual lab or accession effects, are responsible for the formation of the module.

2.2.2 Gene function prediction performance

The co-differential expression networks obtained in the first step of the ENIGMA algorithm were used to assess the gene function prediction performance of the residuals and sample data sets. Topologically,

the residuals network and the sample networks contain comparable numbers of genes and co-differential expression links (edges) and a similar network density and clustering coefficient (Table 2.2). Forty-eight genes in the residuals network are not observed in any sample network, but there is no obvious functional theme among them. For most Gene Ontology (GO)[92] categories, the residuals network contains similar numbers of annotated nodes as the average sample network (Supplemental Data Set 5), but the residuals network contains a significantly higher fraction of genes which are not annotated in the GO database (Table 2.2). Well-represented categories in the residuals network (relative to the sample networks) include categories related to secondary and lipid metabolism, cell wall biogenesis, and pollination. Several photosynthesis- and amino acid metabolism-related categories are relatively poorly represented (Supplemental Data Set 5).

Table 2.2: Topological parameters for the residuals and sample co-differential expression networks. For the sample networks, mean values $\pm$ 1 standard deviation are indicated. Approximate P values are based on the rank of the residuals network relative to the 1000 sample networks.

Topological Parameters	Residuals Network	Sample Networks	P Value
Number of nodes	11474	10695 $\pm$ 2606	0.409
Number of edges	165455	152017 $\pm$ 156476	0.314
Network density	0.0025	0.0021 $\pm$ 0.0012	0.240
Clustering coefficient	0.2388	0.2111 $\pm$ 0.0371	0.211
Unannotated gene fraction	0.2210	0.1841 $\pm$ 0.0139	0.005

The presence of a particular gene or biological process in a network does not automatically indicate that the network provides biologically relevant connections for that gene/process. To evaluate the function prediction performance of the residuals and sample networks, we predicted the function of all genes based on the function of their network neighbours and used the available GO annotations as a gold standard to score precision (proportion of predictions that are true positives) and recall (proportion of known annotations recovered by the predictions) for each network over the prediction False Discovery Rate (FDR) threshold range 10e-2 to 10e-11 (see Section 2.4). The F-measure (harmonic mean of precision and recall) was used as a single integrated measure of prediction performance. An unavoidable pitfall in this approach is the occurrence of false positive and false negative functional annotations in the GO reference set, undermining its use as a gold standard. Although the calculated precision and recall values may therefore deviate from the real values, our approach is still useful for comparative purposes, since similar biases presumably exist for all networks. If any differential bias would exist, one may be inclined to think it might be a bias favouring the sample networks, since comparatively more of the existing GO annotations and supporting experimental evidence can be assumed to derive from major effect perturbations on pooled plant samples, as in the sample data sets, than from minor effect perturbations on individual plants, as in the residuals data set. The fact that significantly more functionally non-annotated genes are recovered in the residuals network than in the average sample network (Table 2.2) may point in this direction, but this can hardly be taken as solid evidence for a differential bias.

Overall, the residuals network produces slightly more predictions for slightly more genes than the average sample network at each FDR threshold (Figure 2.5). For more stringent FDR thresholds, the resulting number of predictions per predicted gene is substantially larger for the residuals network than for the average sample network, reaching the 90th sample networks percentile at FDR = 10e-11. The prediction performance of all networks was assessed for a wide range of GO categories (Figure

2.6), which were classified in five performance categories depending on their F-measure scores for the residuals network relative to the sample networks over the entire prediction FDR range (see Section 2.4). Performance plots for some representative GO categories are depicted in Figure 2.7 (Supplemental Data Set 6 and Figure 2.6 for other categories). The residuals network outperforms the majority of the sample networks for functional categories such as response to wounding, defense response, response to fungus, drought and salt stress responses, response to jasmonic acid (JA), abscisic acid (ABA) and ethylene stimulus, cell communication, lipid and carbohydrate metabolism, and leaf development. On the other hand, the residuals network scores comparatively worse for categories such as responses to light intensity, desiccation, insect, virus, UV and DNA damage, photosynthesis, responses to auxin and

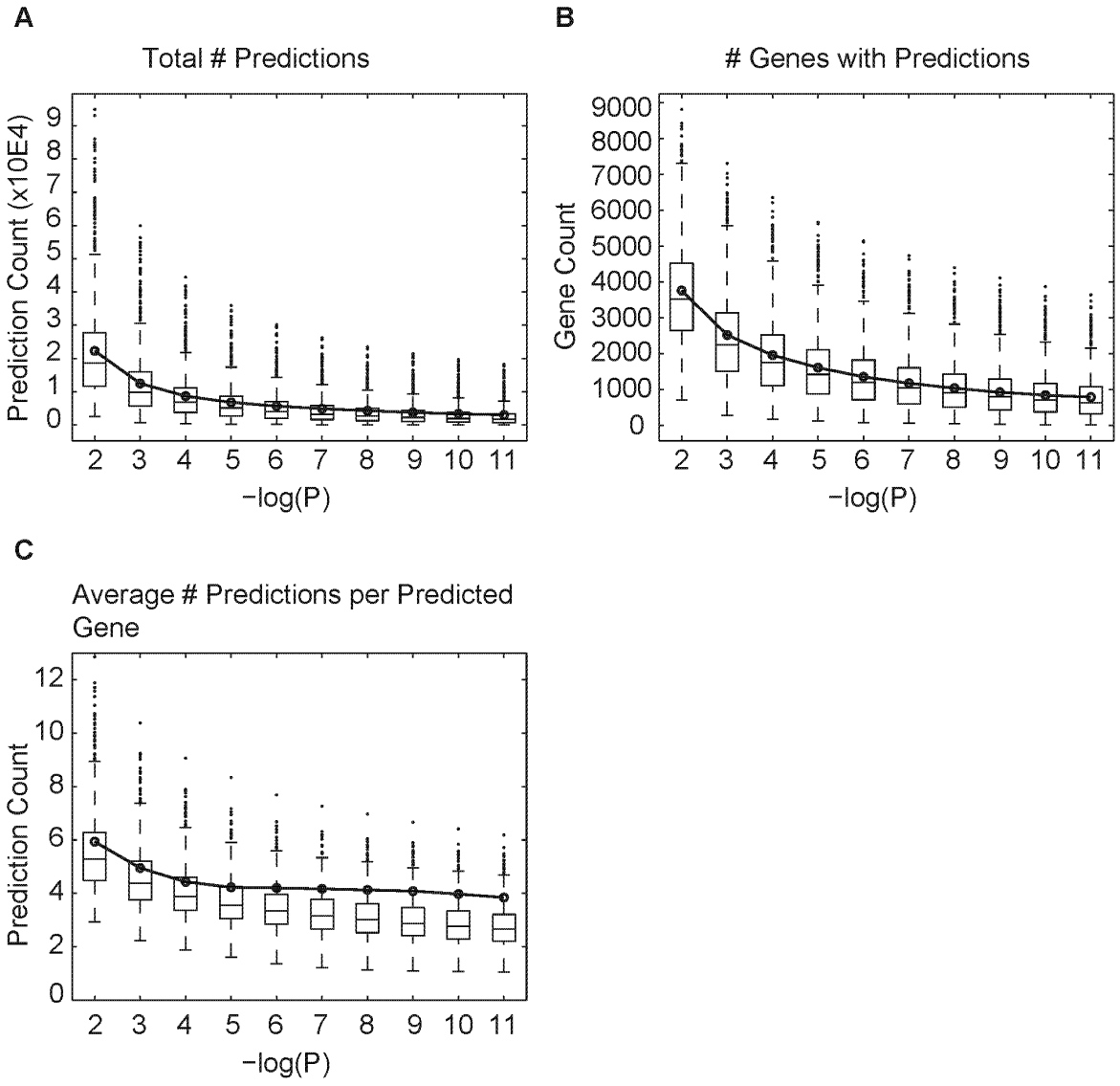


Figure 2.5: Functional prediction statistics for the residuals and sample data sets. (**A.**) Total number of functional predictions per data set at FDR thresholds in the range [10e-02 10e-11]. Only max-depth functional predictions are taken into account, i.e., all predictions for functional categories that are parents or higher-level ancestors of other predicted categories in the GO hierarchy were pruned. (**B.**) Numbers of genes for which functional predictions were made at particular FDR thresholds, for each data set. (**C.**) Average number of pruned predictions per gene for different data sets. For each FDR threshold, only the number of genes for which there are functional predictions at that threshold is taken into account. In all panels, the box-and-whisker plot summarises the 1000 sample data sets, and the solid black line depicts the residuals data set. Boxes extend from the 25th to the 75th percentile, with the median indicated by the central black line. Whiskers extend from each end of the box to the most extreme values within 1.5 times the interquartile range from the respective end. Data points beyond this range are displayed as filled black circles.

brassinosteroids, cell cycle, cell differentiation, tropic responses and root and flower development. Other categories such as oxidative stress, temperature and starvation responses, response to bacteria, salicylic acid-mediated signaling, translation and secondary metabolism score average. A noticeable trend for many GO categories is that for more stringent FDR thresholds, the function prediction performance of the residuals network increasingly improves relative to that of the sample networks (see, e.g., "response to mechanical stimulus" in Figure 2.7).

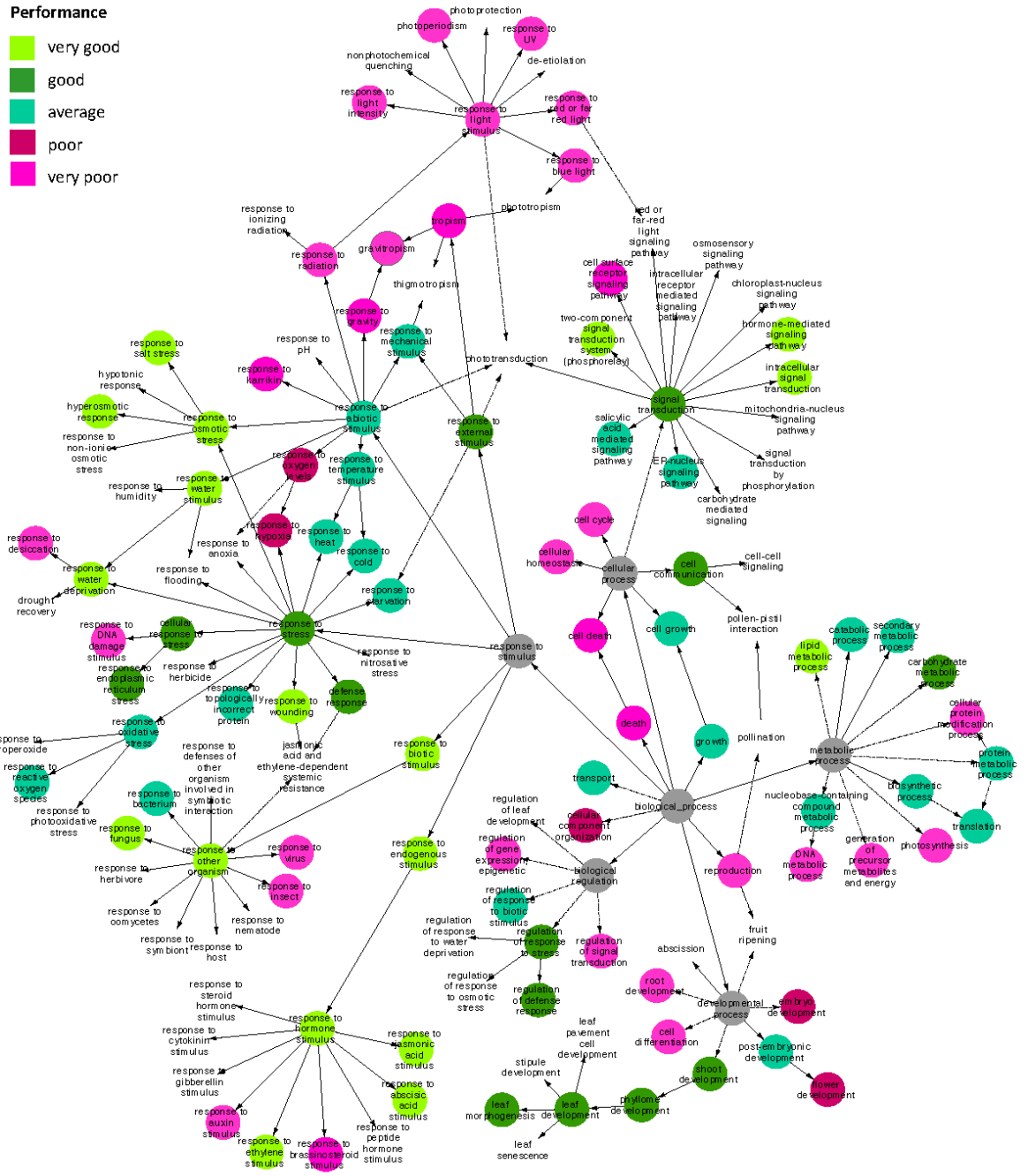


Figure 2.6: Category-specific function prediction performance in the context of the GO hierarchy. Solid arrows represent direct GO relationships, and dash-dot arrows represent indirect relationships. To avoid overcrowding the figure, indirect relationships of GO terms to general functional classes such as 'biological process' (grey nodes) have been omitted if terms have more direct relationships to other terms on the figure. White nodes indicate GO categories for which there is insufficient information to score the performance of the residuals network versus the sample networks. In practice, categories were not scored if less than 500 out of 1000 sample networks gave rise to any predictions at FDR = 10e-2.

Next to the process-centric performance assessment described above, we used a gene-centric method to score the overall gene function prediction performance of all networks (see Section 2.4 and Figure 2.8). Recall values for the residuals network are situated around the 50th percentile of the sample networks over the entire FDR range, but precision scores generally stay below the 25th percentile. The lower precision values of the residuals network with respect to the sample networks may be taken to indicate a genuinely larger amount of false positive gene function predictions. Alternatively, given the incompleteness of the *Arabidopsis* GO annotation[94], it could conceivably be caused by the positive

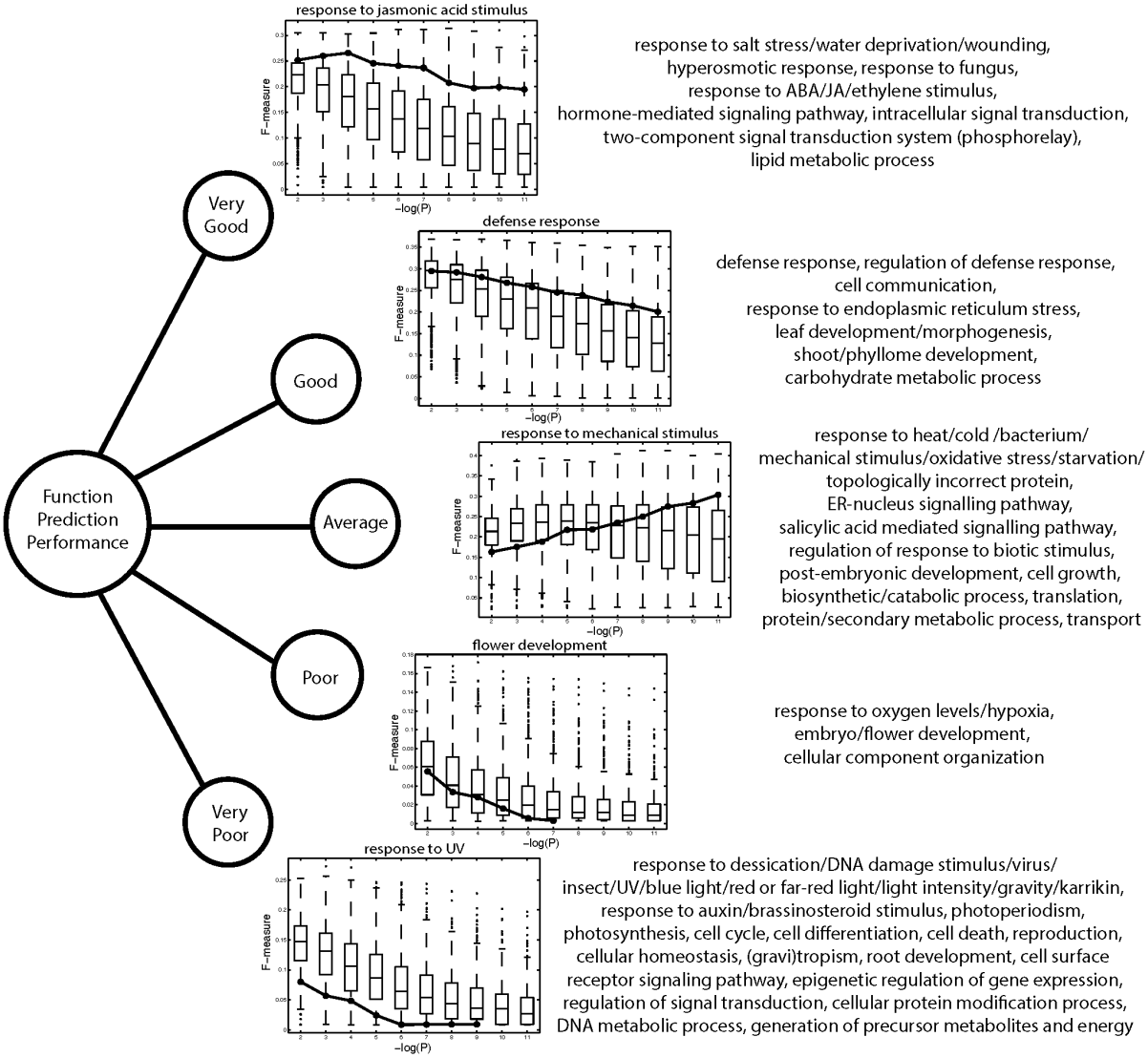


Figure 2.7: Process-specific function prediction performance. Biological processes were subdivided into five performance categories based on the average deviation of the residuals network F-measure from the 25th, 50th and 75th percentiles of the sample network F-measures over the entire FDR range (very good = above the 75th percentile on average; good = on average between the 50th and 75th percentile but closer to the 75th percentile; average = closest to the 50th percentile on average; poor = on average between the 25th and 50th percentile but closer to the 25th percentile; very poor = below the 25th percentile on average, see Section 2.4). An F-measure versus -log(P) (FDR threshold) plot is shown for one representative process per category. Box-and-whisker plots indicate the F-measure distribution over all 1000 sample networks at any given FDR threshold, and the solid line depicts the F-measure trend for the residuals network. Boxes extend from the 25th to the 75th percentile, with the median indicated by the central black line. Whiskers extend from each end of the box to the most extreme values within 1.5 times the interquartile range from the respective end. Data points beyond this range are displayed as little black circles. The categorisation of other processes is shown on the right (Supplemental Data Set 6 for performance plots and Figure 2.6 for a depiction of the tested categories in their GO context). Categories related to environmental stress factors that cannot easily be homogenised across plants generally score above average, as well as the corresponding hormonal responses, while categories related to stresses that are largely absent under lab growth conditions score below average.

identification of a larger amount of false negative functional annotations in the GO reference set, in particular if, as hypothesised above, there were a bias of known GO annotations towards predictions made by the sample data sets, which remains to be proven. As a result of the lower precision values, the global gene function prediction performance of the residuals network at FDR = 10e-2 scores below the 27th percentile of the sample networks (Figure 2.8), but as was the case for many individual GO categories, the residuals network performance increases relative to that of the sample networks for more stringent FDR thresholds, culminating in an F-measure equal to the 55th sample network percentile for FDR = 10e-11. A relative increase of the residuals performance with respect to the sample networks for more stringent FDR thresholds may be expected if there were a bias of the existing GO annotations towards the sample data set predictions. In that case, one would expect a more fair performance balance between the residuals and sample networks for the most confident predictions (which are arguably the most likely to be recovered from any data set), and an increasing bias for predictions at the higher end of

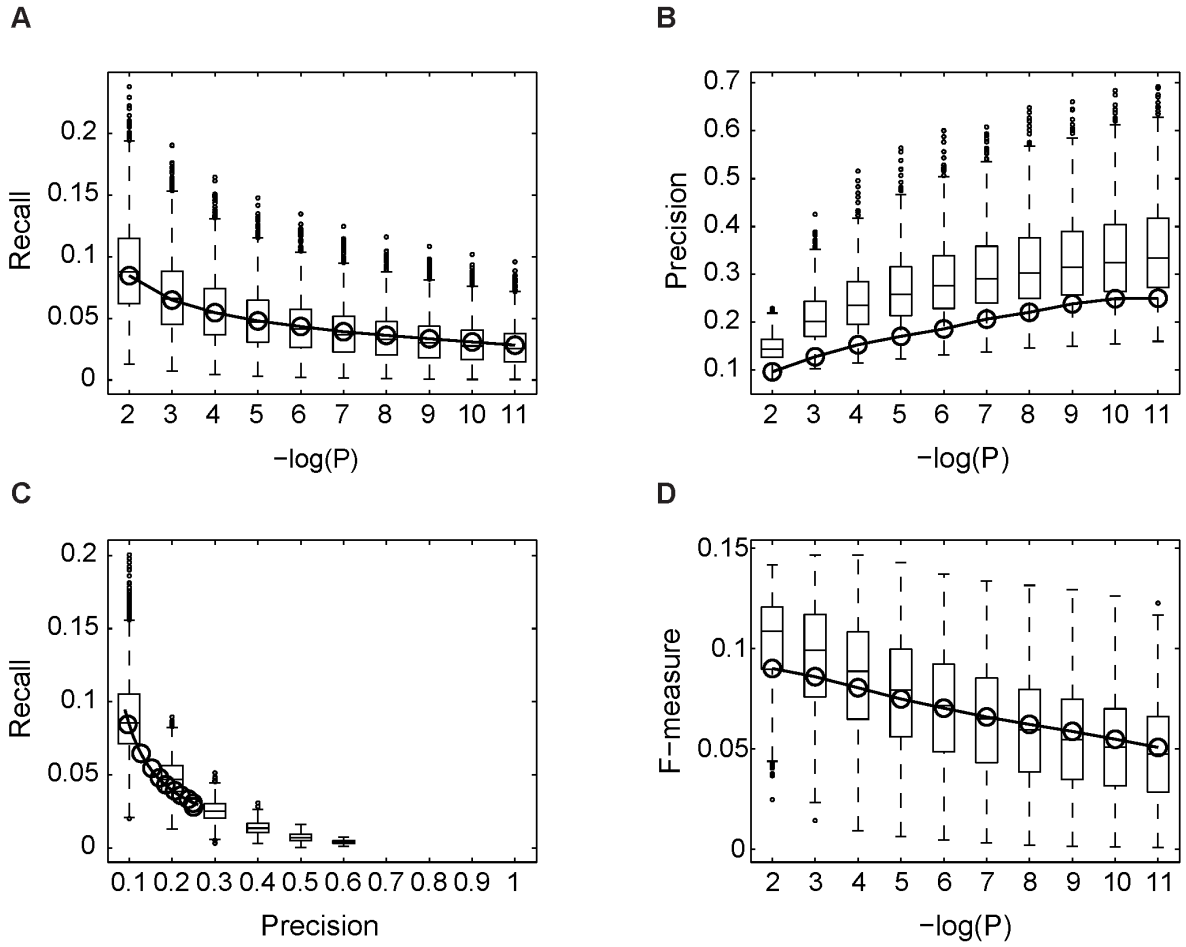


Figure 2.8: Global function prediction performance. Plots **A** to **D** depict the performance of the residuals network (open circles and solid line) and the sample networks (box-and-whisker plots) based on the use of a gene-centric method[93] to score the recall and precision of function predictions across all genes in a given network. Boxes extend from the 25th to the 75th percentile, with the median indicated by the central black line. Whiskers extend from each end of the box to the most extreme values within 1.5 times the interquartile range from the respective end. Data points beyond this range are displayed as little black circles. (**A.**) Recall as a function of the prediction FDR threshold; (**B.**) Precision versus prediction FDR threshold; (**C.**) Precision-recall curve; (**D.**) F-measure as a function of the FDR threshold. Whereas the recall values for the residuals network are situated around the 50th percentile of the sample networks, precision values are generally below the 25th percentile. The combined F-measure score of the residuals network ranges from the 27th sample network percentile for FDR=10e-2 to the 55th percentile for FDR=10e-11.

the FDR range, as observed in Figure 2.8. But again, despite being suggestive, this can hardly be taken as solid evidence for the existence of any bias.

2.2.3 JA signalling case study

Response to JA stimulus (GO:0009753) is one of the best scoring functional categories in the functional prediction performance assessment described above. To assess whether the residuals data set can be used to successfully predict the involvement of novel genes in this process, we screened all networks for novel candidate genes that are a priori annotated as biological regulators (GO:0065007) but are not known to be involved in the JA signaling response (see Section 2.4). *ILL6* came out as the top predicted novel candidate regulator in the residuals network (P = 3.33e-09), with a substantial lead over other candidate genes (Table 2.3). The *ILL6* prediction was supported by 598 out of 1000 sample networks and ranked as the top prediction in 285 of those networks. At least one other computational study also predicted *ILL6* to be involved in the response to JA stimulus[98], but hard experimental evidence has been lacking until now.

Table 2.3: Regulatory genes (GO:0065007) predicted to be involved in the response to jasmonic acid stimulus (GO:0009753) based on the residuals co-differential expression network, at FDR = 0.01. The last column gives the number of sample networks that support the residuals prediction at FDR = 0.01. The involvement of the top-scoring gene *ILL6* in the JA signaling response was validated in this study. We screened literature for direct or indirect evidence supporting the top-10 predictions. Predictions supported by direct evidence in literature are highlighted in green, while predictions supported indirectly (evidence for involvement in related processes or direct evidence for homologs in other species) are highlighted in yellow. Relevant references are indicated by superscripts in the second column.

ORF	Name	Description	Corrected P Value	# Sample Network Predictions
AT1G44350	ILL6	IAA-LEUCINE RESISTANT (ILR)-LIKE GENE 6	3.33E-09	598
AT2G14750	APK[95]	APS KINASE	1.52E-04	88
AT1G52890	ANAC019[96]	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 19	1.52E-04	152
AT1G28370	ERF11	ERF DOMAIN PROTEIN 11	1.52E-04	257
AT2G47190	MYB2	MYB DOMAIN PROTEIN 2	3.98E-04	38
AT3G61190	BAP1[97]	BON ASSOCIATION PROTEIN 1	0.001484	168
AT2G36080	AT2G36080	B3 DOMAIN-CONTAINING PROTEIN	0.001505	13
AT4G17500	ATERF-1	ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 1	0.001505	342
AT5G57150	AT5G57150	TRANSCRIPTION FACTOR BHLH35	0.001601	13
AT1G19270	DA1	DA 1	0.002700	13
AT4G28560	RIC7	ROP-INTERACTIVE CRIB MOTIF-CONTAINING PROTEIN 7	0.003085	0
AT2G40140	SZF2	(SALT-INDUCIBLE ZINC FINGER 2	0.003085	318
AT1G42990	BZIP60	BASIC REGION/LEUCINE ZIPPER MOTIF 60	0.003383	48
AT3G08720	S6K2	ARABIDOPSIS THALIANA SERINE/THREONINE PROTEIN KINASE 2	0.004279	221
AT4G23700	CHX17	CATION/H+ EXCHANGER 17	0.004304	34
AT3G26830	PAD3	PHYTOALEXIN DEFICIENT 3	0.005540	90
AT3G02875	ILR1	IAA-LEUCINE RESISTANT 1	0.005540	48
AT5G67450	AZF1	ARABIDOPSIS ZINC-FINGER PROTEIN 1	0.005679	10
AT3G50750	BEH1	BES1/BZR1 HOMOLOG 1	0.005679	4
AT5G06870	PGIP2	POLYGALACTURONASE INHIBITING PROTEIN 2	0.005907	173
AT3G02380	COL2	CONSTANS-LIKE 2	0.006495	149
AT2G18950	HPT1	HOMOGENTISATE PHYTYLTRANSFERASE 1	0.006730	6
AT3G23010	ATRLP36	RECEPTOR LIKE PROTEIN 36	0.007067	0
AT4G37180	AT4G37180	MYB FAMILY TRANSCRIPTION FACTOR	0.007547	7
AT2G38170	RCI4	RARE COLD INDUCIBLE 4	0.009002	5
AT2G44840	ERF13	ETHYLENE-RESPONSIVE ELEMENT BINDING FACTOR 13	0.009002	193
AT5G54310	AGD5	ARF-GAP DOMAIN 5	0.009002	6
AT5G26920	CBP60G	CAM-BINDING PROTEIN 60-LIKE G	0.009002	150
AT2G41900	AT2G41900	ZINC FINGER CCCH DOMAIN-CONTAINING PROTEIN 30	0.009002	3
AT1G61800	GPT2	GLUCOSE-6-PHOSPHATE/PHOSPHATE TRANSLOCATOR 2	0.009002	21
AT5G27520	PNC2	PEROXISOMAL ADENINE NUCLEOTIDE CARRIER 2	0.009002	87

We took a reverse-genetics approach to investigate the possible role of *ILL6* in jasmonate signaling. Two homozygous T-DNA insertional mutant lines, *ill6-1* and *ill6-2*, were identified in which no full-length transcript of *ILL6* was detectable by RT-PCR (Figure 2.10). To examine the mutant's sensitivity to the hormone, these plant lines and the wild-type, Columbia-0 (Col-0), were grown on various concentrations of methyl jasmonate (MeJA), and the root lengths and shoot weights were determined (Figure 2.9-A and 2.9-B). Analysis of these data indicate that the roots of *ill6-1* and *ill6-2* are significantly shorter and the rosettes weigh significantly less than those of the wild type across all levels of MeJA treatment ($P = 0.0011$ and $P < 0.0001$, respectively; see Section 2.4 for details on statistical analyses). There is also a slight but significant ($P = 0.0298$) genotype $\times$ MeJA treatment effect in terms of shoot weight response to MeJA. Thus, the mutants are slightly but significantly more sensitive to exogenous jasmonate than the wild type. Furthermore, liquid chromatography-tandem mass spectrometry analysis revealed that the two mutants both accumulate substantially more wound-induced jasmonoyl-Ile (JA-Ile) than the wild type (Figure 2.9-C; $P = 0.0001$ for the genotype effect and $P = 0.0003$ for the genotype $\times$ time interaction effect). Together, these data are consistent with ILL6 acting as a negative regulator of the jasmonate response. It is an attractive hypothesis that ILL6 could be a JA-Ile hydrolase, cleaving the JA-Ile amide bond in vivo and releasing Ile and molecularly inactive JA. ILL6 is a member of a family of proteins whose founding member, ILR1, has been characterised as an auxin-Leu hydrolase[99], while a second member,

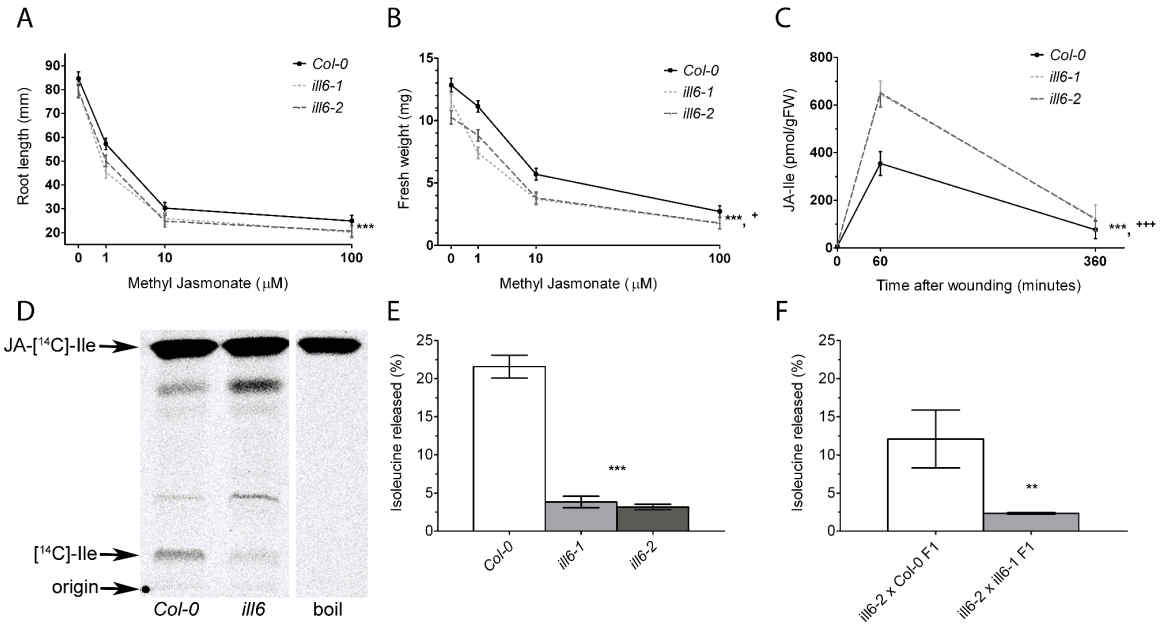


Figure 2.9: *ILL6* Negatively Regulates JA Response and Wound-Induced JA-Ile Accumulation, Likely through Hydrolysis of JA-Ile. (**A.**) Response of mutant and wild-type *Arabidopsis* seedling's root length to exogenous MeJA. Seedlings were exposed to media containing 0, 1, 10, or 100 μM MeJA for 8 d ($n \geq 16$ seedlings). (**B.**) The rosettes of the plants in **(A)** were excised from the roots and weighed ($n \geq 16$ seedlings). (**C.**) Time course of wound-induced JA-Ile accumulation. Plants were wounded and damaged leaves were harvested at the indicated time points after wounding and JA-Ile accumulation was analysed by liquid chromatography-tandem mass spectroscopy (n=6 plants across two independent experiments). (**D.**) Representative in vivo JA-[^{14}C]-Ile hydrolysis assay. JA-[^{14}C]-Ile was applied to individual plant leaves of the indicated genotype and extracts were separated by thin-layer chromatography and visualised by autoradiography. (**E.**) In vivo hydrolysis of JA-[^{14}C]-Ile in *ill6* mutants and the wild type. Autoradiograms were quantified by densitometry ($n \geq 9$ plants across five independent experiments). (**F.**) *ill6-1* and *ill6-2* are allelic mutations. The two F1 hybrids indicated were subjected to an in vivo hydrolysis assay as in **D** and **E** ($n = 3$ plants). For all plots, data represent mean $\pm$ SE, asterisks indicate significance of genotype effects: $*$, $p \leq 0.05$, $**$, $p \leq 0.01$, $***$ $p \leq 0.001$. + indicates $p \leq 0.05$ for the genotype $\times$ MeJA interaction effect in panel **B**, +++ indicates $p \leq 0.001$ for the genotype $\times$ time interaction effect in panel **C** (see Section 2.4 for details on statistical analyses).

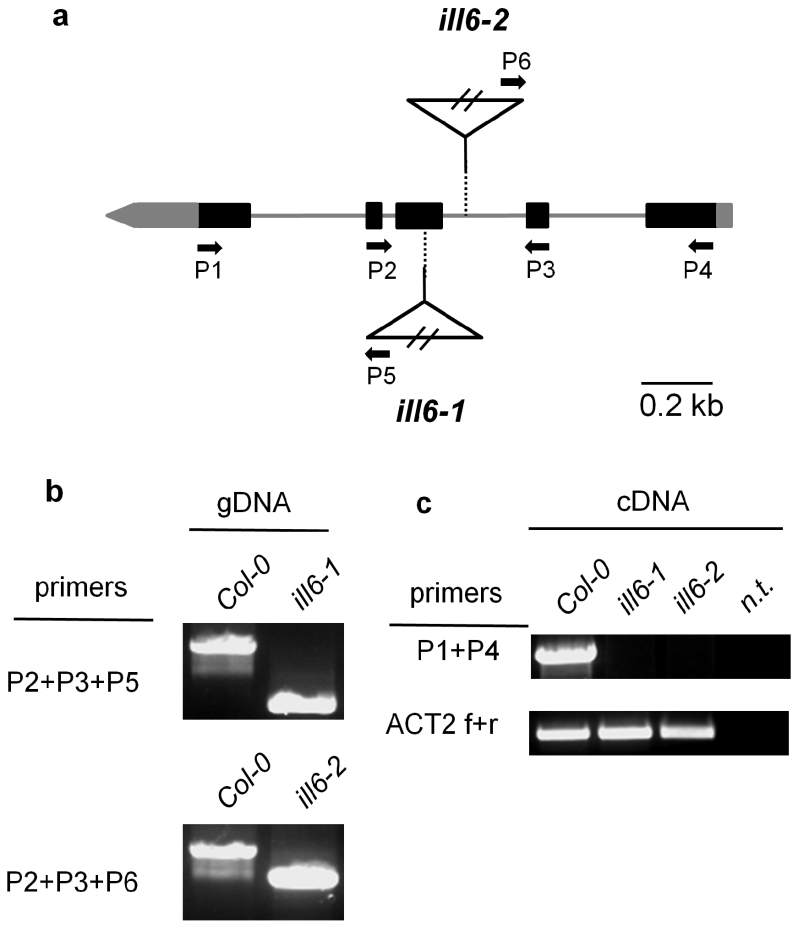


Figure 2.10: Identification of *ill6* mutants. (a) Diagram of the *ILL6* locus, showing intron (lines) and exon (black boxes) structure, T-DNA insertion sites of the *ill6-1* and *ill6-2* alleles, and sites of oligonucleotide primers used for PCR analysis. **(b)** Identification of homozygous mutants by PCR analysis of genomic DNA. **(c)** RT-PCR analysis of *ILL6* transcript accumulation in Col-0 and the *ill6-1* and *ill6-2* mutants (40 cycles). Amplification of the *ACTIN2* transcript served as a control (30 cycles). (n.t., no template). PCR experiments were performed three times with similar results.

IAR3, is known to be an auxin-Ala hydrolase in Arabidopsis[100]. Furthermore, the IAR3 homolog from *Nicotiana attenuata*, Na IAR3, was recently shown to be a JA-Ile hydrolysing enzyme[101]. We expressed a recombinant ILL6 protein in *Escherichia coli*, but to date we have not detected any JA-Ile hydrolase activity from this protein, nor have we seen in vitro activity on several other tested JA-amino acid conjugates.

To address the in vivo activity of this protein, we examined the metabolic fate of exogenously applied radiolabeled JA-Ile (see Section 2.4). JA-[^{14}C]Ile was applied to individual leaves of wild- type and *ill6* mutant plants, and after 24 h, ethanolic extracts of these treated leaves were separated by thin layer chromatography (Figure 2.9-D). Autoradiographic detection revealed that whereas boiled leaf controls produced no detectable radiolabeled metabolic products of JA-[^{14}C]Ile, ;20% of the radioactivity applied to the wild-type Col-0 was released as free [^{14}C]Ile. This result was in marked contrast to either *ill6* mutant, in which only 4% of applied radioactivity was released as [^{14}C]Ile (Figure 2.9-E; $\log_{10}$-transformed one-way ANOVA *F*-test $P < 0.0001$). Next, for a complementation test, we crossed *ill6-2* as the pollen donor to both Col-0 and *ill6-1*. In the F1 hybrids between the mutant and wild type, we observed a release of 12% of applied radioactivity as [^{14}C]Ile, whereas in the F1 hybrids between the two mutants, we observed little release of [^{14}C]Ile, similar to both mutant parents (Figure 2.9-F; *t* test on $\log_{10}$-transformed data, P =

0.0067). This complementation test thus indicates that the biochemical defect in JA-[^{14}C]Ile hydrolysis is due to the *ill6* mutant lesions. Collectively from these data, we conclude that ILL6 is a negative regulator of jasmonate accumulation and response, likely through its role as an amidohydrolase of JA-Ile, though formally we cannot exclude the possibility that ILL6 acts on an in planta-produced derivative of JA-Ile.

2.2.4 Literature screen for direct and indirect evidence supporting the top 10 residuals predictions for various GO categories in the "Very Good" performance class

Above, we provide evidence validating the functional prediction of a gene (*ILL6*) that is also predicted to be involved in the JA signaling response by the majority of sample networks and that as such cannot be regarded as a prediction that is unique to the residuals network. In fact, for most of the categories we screened, there are barely any residuals predictions that are not supported by at least one sample data set (e.g., there are only two such predictions out of 31, for "response to JA stimulus"; see Section 2.3), showing that the residuals data set does generally not make predictions that are beyond the reach of any other data set. Although high-confidence residuals predictions that are made by a higher number of randomly sampled compendia, such as the ILL6 prediction, may to some extent be viewed as being more supported and may be prioritised as such for wet-lab testing, residuals predictions that are rarely recovered by the sample data sets may, if validated, point to specific advantages of profiling uncontrolled expression variation across individuals.

To investigate the added value of profiling expression responses to micro-environmental variability among individuals in more detail, we screened literature for direct or indirect evidence supporting the top 10 novel predictions for six GO categories that were classified in the "very good" prediction performance category, namely, the response to JA, ABA, and ethylene stimulus, response to fungus, response to salt stress, and response to water deprivation. Although literature screens can arguably never be all-encompassing, we did find reports describing direct (indirect) experimental evidence for one (two) JA predictions, three (one) ABA predictions, one (two) ethylene predictions, two (two) response to fungus predictions, one (0) response to salt stress predictions, and 0 (one) response to water deprivation predictions out of the top 10 for each category (Tables 2.3 to 2.8). As for the direct evidence, these are essentially earlier findings that have not yet been incorporated in the GO database, and our associated predictions can as such not really be regarded as novel, although supported. The indirect evidence references given in Tables 2.3 to 2.8 link the predicted gene to a process or pathway related to the target process or describe direct evidence for a homolog of the predicted gene in another species. Although more than half (10/16) of the top 10 residuals predictions for which we recovered supporting experimental evidence in literature are also predicted by a sizeable proportion of sample networks (14.6 to 41.7%), we did find a substantial number (6/16) of supported residuals predictions that are only predicted by <10% of the sample networks, in particular among the indirectly supported predictions (4/8, as opposed to 2/8 for directly supported predictions). Directly supported residuals predictions that are uncovered by <10% of the sample networks include the involvement of AZF1 (3.8%) in the response to ABA stimulus[102] and TGA5 (0.3%) in ethylene signaling[103] (numbers in parentheses indicate sample network prediction percentages). Indirectly supported residuals predictions include the involvement of APK1 (8.8%) in JA

signaling, MPK1 (0.7%) and JAZ1 (7.0%) in ethylene signaling, and CRT3 (1.6%) in the response to water deprivation. APK1 was previously reported to be involved in the synthesis of glucosinolates and sulfated 12-hydroxyjasmonate[95]. MPK1 activity was shown earlier to be repressed by the ethylene response regulator CTR1, but the physiological relevance of MPK1 downregulation for ethylene signaling responses is still unclear[104]. JAZ1 was reported to interact with and repress the ethylene-stabilised transcription factors EIN3 and EIL1[105]. And a putative ortholog of CRT3 in wheat (*Triticum aestivum*) was previously shown to be involved in drought stress response[106]. Although we do not claim that the residuals data set is superior for all functional prediction purposes, these results suggest that the residuals data set can produce valid novel predictions that are seldom recovered from randomly sampled perturbational data sets.

Table 2.4: Regulatory genes (GO:0065007) predicted to be involved in the response to abscisic acid stimulus (GO:0009737) based on the residuals co-differential expression network, at FDR = 0.01. The last column gives the number of sample networks that support the residuals prediction at FDR = 0.01. We screened literature for direct or indirect evidence supporting the top-10 predictions. Predictions supported by direct evidence in literature are highlighted in green, while predictions supported indirectly (evidence for involvement in related processes or direct evidence for homologs in other species) are highlighted in yellow. Relevant references are indicated by superscripts in the second column.

ORF	Name	Description	Corrected P Value	# Sample Network Predictions
AT5G26920	CBP60G [107]	CAM-BINDING PROTEIN 60-LIKE G	9.76E-07	323
AT5G67450	AZF1 [102]	ARABIDOPSIS ZINC-FINGER PROTEIN 1	7.86E-05	38
AT1G09530	PIF3	PHYTOCHROME-INTERACTING FACTOR 3	7.86E-05	34
AT3G08720	S6K2	ARABIDOPSIS THALIANA SERINE/THREONINE PROTEIN KINASE 2	2.63E-04	367
AT5G47220	ERF2	ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 2	4.73E-04	96
AT2G17040	ANAC036	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 36	4.73E-04	96
AT4G31800	WRKY18 [108]	WRKY TRANSCRIPTION FACTOR 18	4.73E-04	149
AT5G14960	E2L1	DP-E2F-LIKE 2	5.83E-04	8
AT1G51660	MKK4 [109]	MITOGEN-ACTIVATED PROTEIN KINASE KINASE 4	6.41E-04	146
AT1G73260	KTI1	KUNITZ TRYPSIN INHIBITOR 1	6.43E-04	22
AT3G52400	SYP122	SYNTAXIN OF PLANTS 122	6.43E-04	453
AT5G13080	WRKY75	PUTATIVE WRKY TRANSCRIPTION FACTOR 75	6.43E-04	99
AT5G39610	ATNAC2	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 2	7.45E-04	83
AT4G17500	ATERF-1	ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 1	8.82E-04	237
AT1G29400	AML5	ARABIDOPSIS MEI2-LIKE PROTEIN 5	8.82E-04	58
AT3G07780	OBE1	OBERON1	0.001399	127
AT4G33050	EDA39	EMBRYO SAC DEVELOPMENT ARREST 39	0.001527	482
AT1G72830	HAP2C	NUCLEAR TRANSCRIPTION FACTOR Y SUBUNIT A-3	0.001527	1
AT2G47890	AT2G47890	ZINC FINGER PROTEIN CONSTANS-LIKE 13	0.001527	111
AT1G60190	PUB19	U-BOX DOMAIN-CONTAINING PROTEIN 19	0.001527	447
AT2G41010	CAMBP25	CALMODULIN BINDING PROTEIN 25	0.001924	99
AT2G40140	SZF2	(SALT-INDUCIBLE ZINC FINGER 2	0.003681	566
AT5G47910	RBOHD	RESPIRATORY BURST OXIDASE HOMOLOGUE D	0.003681	97
AT1G80840	WRKY40	PUTATIVE WRKY TRANSCRIPTION FACTOR 40	0.003681	289
AT5G24120	SIGE	SIGMA FACTOR E	0.003681	87
AT1G42990	BZIP60	BASIC REGION/LEUCINE ZIPPER MOTIF 60	0.003896	227
AT3G23010	ATRLP36	RECEPTOR LIKE PROTEIN 36	0.003896	1
AT4G25960	PGP2	P-GLYCOPROTEIN 2	0.003896	39
AT3G52430	PAD4	PHYTOALEXIN DEFICIENT 4	0.003896	165
AT3G55980	SZF1	SALT-INDUCIBLE ZINC FINGER 1	0.003921	413
AT2G30140	AT2G30140	UDP-GLUCORONOSYL/UDP-GLUCOSYL TRANSFERASE-LIKE PROTEIN	0.004650	80
AT1G51140	AT1G51140	TRANSCRIPTION FACTOR BHLH122	0.004652	167
AT1G53170	ERF8	ETHYLENE-RESPONSIVE TRANSCRIPTION FACTOR 8	0.005976	22
AT1G52890	ANAC019	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 19	0.005976	263
AT1G28370	ERF11	ERF DOMAIN PROTEIN 11	0.006149	294
AT4G12720	GFG1	GROWTH FACTOR GENE 1	0.006192	381
AT4G25480	CBF3	C-REPEAT BINDING FACTOR 3	0.006192	175
AT5G58620	AT5G58620	ZINC FINGER CCCH DOMAIN-CONTAINING PROTEIN 66	0.006661	20
AT5G24470	PRR5	PSEUDO-RESPONSE REGULATOR 5	0.007385	71
AT5G27520	PNC2	PEROXISOMAL ADENINE NUCLEOTIDE CARRIER 2	0.007938	105
AT3G61190	BAP1	BON ASSOCIATION PROTEIN 1	0.008043	270
AT4G17230	SCL13	SCARECROW-LIKE 13	0.008043	385
AT3G59700	LECRK1	LECTIN-RECEPTOR KINASE 1	0.008771	181
AT1G22280	PAPP2C	PUTATIVE PROTEIN PHOSPHATASE 2C 9	0.008848	342
AT5G13820	TBP1	TELOMERIC DNA BINDING PROTEIN 1	0.009217	47

Table 2.5: Regulatory genes (GO:0065007) predicted to be involved in the response to ethylene stimulus (GO:0009723) based on the residuals co-differential expression network, at FDR = 0.01. The last column gives the number of sample networks that support the residuals prediction at FDR = 0.01. We screened literature for direct or indirect evidence supporting the top-10 predictions. Predictions supported by direct evidence in literature are highlighted in green, while predictions supported indirectly (evidence for involvement in related processes or direct evidence for homologs in other species) are highlighted in yellow. Relevant references are indicated by superscripts in the second column.

ORF	Name	Description	Corrected P Value	# Sample Network Predictions
AT2G40140	SZF2	SALT-INDUCIBLE ZINC FINGER 2	3.90E-04	223
AT1G10210	MPK1 [104]	MITOGEN-ACTIVATED PROTEIN KINASE 1	0.001006	7
AT1G19180	JAZ1 [105]	JASMONATE-ZIM-DOMAIN PROTEIN 1	0.001006	70
AT2G25900	ATCTH	ZINC FINGER CCCH DOMAIN-CONTAINING PROTEIN 23	0.001006	94
AT5G06960	TGA5 [103]	TGACG MOTIF-BINDING FACTOR 5	0.002658	3
AT3G26520	TIP2	TONOPLAST INTRINSIC PROTEIN 2	0.002812	4
AT5G27520	PNC2	PEROXISOMAL ADENINE NUCLEOTIDE CARRIER 2	0.002812	10
AT5G13680	ABO1	ABA-OVERLY SENSITIVE 1	0.003961	0
AT3G61190	BAP1	BON ASSOCIATION PROTEIN 1	0.004675	159
AT3G26830	PAD3	PHYTOALEXIN DEFICIENT 3	0.005208	32
AT2G25440	ATRLP20	RECEPTOR LIKE PROTEIN 20	0.005222	5
AT5G45110	NPR3	NPR1-LIKE PROTEIN 3	0.005257	62
AT3G50750	BEH1	BES1/BZR1 HOMOLOG 1	0.005672	9
AT5G39660	CDF2	CYCLING DOF FACTOR 2	0.005672	71
AT3G17100	AT3G17100	TRANSCRIPTION FACTOR BHLH147	0.005896	28
AT3G02380	COL2	CONSTANS-LIKE 2	0.006656	155
AT4G31800	WRKY18	WRKY TRANSCRIPTION FACTOR 18	0.008372	43
AT3G19360	AT3G19360	ZINC FINGER CCCH DOMAIN-CONTAINING PROTEIN 39	0.008372	0
AT2G39250	SNZ	SCHNARCHZAPFEN	0.008372	0
AT3G52400	SYP122	SYNTAXIN OF PLANTS 122	0.008521	104
AT2G20180	PIF1	PHY-INTERACTING FACTOR 1	0.009944	14

Table 2.6: Regulatory genes (GO:0065007) predicted to be involved in the response to fungus (GO:0009620) based on the residuals co-differential expression network, at FDR = 0.01. The last column gives the number of sample networks that support the residuals prediction at FDR = 0.01. We screened literature for direct or indirect evidence supporting the top-10 predictions. Predictions supported by direct evidence in literature are highlighted in green, while predictions supported indirectly (evidence for involvement in related processes or direct evidence for homologs in other species) are highlighted in yellow. Relevant references are indicated by superscripts in the second column.

ORF	Name	Description	Corrected P Value	# Sample Network Predictions
AT1G73500	MKK9 [110]	MAP KINASE KINASE 9	1.55E-06	178
AT1G19180	JAZ1 [111]	JASMONATE-ZIM-DOMAIN PROTEIN 1	1.55E-06	272
AT4G17500	ATERF-1 [112]	ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 1	2.67E-04	417
AT1G28370	ERF11	ERF DOMAIN PROTEIN 11	4.07E-04	222
AT3G10500	ANAC053	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 53	4.60E-04	106
AT3G57530	CPK32	CALCIUM-DEPENDENT PROTEIN KINASE 32	4.60E-04	199
AT3G26830	PAD3 [113]	PHYTOALEXIN DEFICIENT 3	6.01E-04	380
AT2G28160	FRU	FER-LIKE REGULATOR OF IRON UPTAKE	7.13E-04	4
AT2G17040	ANAC036	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 36	9.91E-04	200
AT1G25560	TEM1	TEMPRANILLO 1	0.001149	67
AT2G40750	WRKY54	WRKY DNA-BINDING PROTEIN 54	0.001235	274
AT3G07780	OBE1	OBERON1	0.00162	91
AT1G53170	ERF8	ETHYLENE-RESPONSIVE TRANSCRIPTION FACTOR 8	0.001899	2
AT5G13220	JAS1	JASMONATE-ASSOCIATED 1	0.001899	105
AT2G19450	TAG1	DIACYLGLYCEROL O-ACYLTRANSFERASE 1	0.002134	12
AT3G59700	LECRK1	LECTIN-RECEPTOR KINASE 1	0.002344	311
AT3G52430	PAD4	PHYTOALEXIN DEFICIENT 4	0.002452	309
AT1G28480	GRX480	GLUTAREDOXIN-C9	0.002452	347
AT2G30140	AT2G30140	UDP-GLUCORONOSYL/UDP-GLUCOSYL TRANSFERASE-LIKE PROTEIN	0.002683	85
AT4G29810	MK1	MAP KINASE KINASE 1	0.002683	87
AT1G27730	STZ	SALT TOLERANCE ZINC FINGER	0.002984	408
AT3G01080	WRKY58	WRKY DNA-BINDING PROTEIN 58	0.003049	7
AT5G48400	GLR1.2	GLUTAMATE RECEPTOR 1.2	0.003318	57
AT1G02450	NIMIN1	NIM1-INTERACTING 1	0.003318	123
AT5G40170	ATRLP54	RECEPTOR LIKE PROTEIN 54	0.003351	91
AT3G15210	ERF4	ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 4	0.003486	136
AT1G78600	LZF1	LIGHT-REGULATED ZINC FINGER PROTEIN 1	0.003496	12
AT4G29040	RPT2A	REGULATORY PARTICLE AAA-ATPASE 2A	0.003496	5
AT5G47220	ERF2	ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 2	0.004064	203
AT2G46400	WRKY46	PUTATIVE WRKY TRANSCRIPTION FACTOR 46	0.004171	390
AT1G44350	ILL6	IAA-AMINO ACID HYDROLASE ILR1-LIKE 6	0.004477	46
AT5G27520	PNC2	PEROXISOMAL ADENINE NUCLEOTIDE CARRIER 2	0.004481	96
AT1G06160	ORA59	OCTADECANOID-RESPONSIVE ARABIDOPSIS AP2/ERF 59	0.004481	37
AT1G42990	BZIP60	BASIC REGION/LEUCINE ZIPPER MOTIF 60	0.004567	160
AT1G01560	ATMPK11	MITOGEN-ACTIVATED PROTEIN KINASE 11	0.005033	299
AT5G24470	PRR5	PSEUDO-RESPONSE REGULATOR 5	0.005569	19
AT3G56240	CCH	COPPER CHAPERONE	0.005569	6
AT3G45640	MPK3	MITOGEN-ACTIVATED PROTEIN KINASE 3	0.005569	496
AT3G08720	S6K2	ARABIDOPSIS THALIANA SERINE/THREONINE PROTEIN KINASE 2	0.005569	334
AT1G80830	PMIT1	METAL TRANSPORTER NRAMP1	0.00615	16
AT4G17230	SCL13	SCARECROW-LIKE 13	0.006265	312
AT3G55980	SZF1	SALT-INDUCIBLE ZINC FINGER 1	0.007226	480
AT5G59450	AT5G59450	SCARECROW-LIKE PROTEIN 11	0.007226	32
AT3G02875	ILR1	IAA-LEUCINE RESISTANT 1	0.007226	36
AT1G32640	MYC2	TRANSCRIPTION FACTOR MYC2	0.007254	154
AT4G33050	EDA39	EMBRYO SAC DEVELOPMENT ARREST 39	0.007404	611
AT2G42540	COR15	COLD-REGULATED PROTEIN 15A	0.007418	509
AT5G48380	BIR1	BAK1-INTERACTING RECEPTOR-LIKE KINASE 1	0.008346	520
AT2G43350	ATGPX3	GLUTATHIONE PEROXIDASE 3	0.008805	3
AT1G73260	KTI1	KUNITZ TRYPSIN INHIBITOR 1	0.008805	306
AT4G37180	AT4G37180	MYB FAMILY TRANSCRIPTION FACTOR	0.009021	5
AT3G15500	ANAC055	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 55	0.009814	38

Table 2.7: Regulatory genes (GO:0065007) predicted to be involved in the response to salt stress (GO:0009651) based on the residuals co-differential expression network, at FDR = 0.01. The last column gives the number of sample networks that support the residuals prediction at FDR = 0.01. We screened literature for direct or indirect evidence supporting the top-10 predictions. Predictions supported by direct evidence in literature are highlighted in green. Relevant references are indicated by superscripts in the second column.

ORF	Name	Description	Corrected P Value	# Sample Network Predictions
AT2G46680	ATHB7[114]	ARABIDOPSIS THALIANA HOMEOBOX 7	8.42E-05	264
AT5G13080	WRKY75	PUTATIVE WRKY TRANSCRIPTION FACTOR 75	1.06E-04	56
AT1G09530	POC1	PHOTOCURRENT 1	1.60E-04	16
AT5G47220	ERF2	ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 2	2.63E-04	64
AT2G25110	SDF2	STROMAL CELL-DERIVED FACTOR 2-LIKE PROTEIN PRECURSOR	6.05E-04	669
AT2G43350	ATGPX3	GLUTATHIONE PEROXIDASE 3	8.55E-04	45
AT5G47120	BI-1	BAX INHIBITOR 1	0.001185	37
AT5G27520	PNC2	PEROXISOMAL ADENINE NUCLEOTIDE CARRIER 2	0.001205	89
AT1G28370	ERF11	ERF DOMAIN PROTEIN 11	0.001549	184
AT2G04450	ATNUDT6	ARABIDOPSIS THALIANA NUDIX HYDROLASE HOMOLOG 6	0.001549	0
AT4G34390	XLG2	EXTRA-LARGE GTP-BINDING PROTEIN 2	0.001549	135
AT5G54310	AGD5	ARF-GAP DOMAIN 5	0.002105	15
AT1G10210	MPK1	MITOGEN-ACTIVATED PROTEIN KINASE 1	0.002105	12
AT3G11820	PEN1	PENETRATION1	0.002105	227
AT1G08450	CRT3	CALRETICULIN 3	0.002105	88
AT1G64810	APO1	ACCUMULATION OF PHOTOSYSTEM ONE 1	0.002105	15
AT3G62600	ATERDJ3B	DNAJ HEAT SHOCK FAMILY PROTEIN	0.003377	810
AT3G17100	AT3G17100	TRANSCRIPTION FACTOR BHLH147	0.004021	5
AT1G71220	EBS1	EMS-MUTAGENIZED BRI1 SUPPRESSOR 1	0.004807	279
AT3G53230	AT3G53230	CELL DIVISION CONTROL PROTEIN 48-D	0.004977	94
AT1G14360	UTR3	UDP-GALACTOSE TRANSPORTER 3	0.004977	446
AT3G56290	AT3G56290	HYPOTHETICAL PROTEIN	0.004977	4
AT5G24470	PRR5	PSEUDO-RESPONSE REGULATOR 5	0.004977	31
AT4G17500	ATERF-1	ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 1	0.004977	107
AT4G29040	RPT2A	REGULATORY PARTICLE AAA-ATPASE 2A	0.004977	190
AT1G52890	ANAC019	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 19	0.004977	135
AT3G23010	ATRLP36	RECEPTOR LIKE PROTEIN 36	0.005001	0
AT4G25960	PGP2	P-GLYCOPROTEIN 2	0.005001	55
AT2G02810	UTR1	UDP-GALACTOSE TRANSPORTER 1	0.005001	309
AT5G13820	TBP1	TELOMERIC DNA BINDING PROTEIN 1	0.005026	45
AT3G12250	TGA6	TGACG MOTIF-BINDING FACTOR 6	0.005329	2
AT2G42890	AML2	ARABIDOPSIS-MEI2-LIKE 2	0.005823	21
AT1G19270	DA1	DA 1	0.006114	14
AT1G19180	JAZ1	JASMONATE-ZIM-DOMAIN PROTEIN 1	0.006114	117
AT3G26830	PAD3	PHYTOALEXIN DEFICIENT 3	0.006114	22
AT3G02875	ILR1	IAA-LEUCINE RESISTANT 1	0.006159	54
AT4G27410	RD26	RESPONSIVE TO DESICCATION 26	0.006368	198
AT1G72830	HAP2C	NUCLEAR TRANSCRIPTION FACTOR Y SUBUNIT A-3	0.006368	5
AT2G30140	AT2G30140	UDP-GLUCORONOSYL/UDP-GLUCOSYL TRANSFERASE-LIKE PROTEIN	0.006368	119
AT1G61800	GPT2	GLUCOSE-6-PHOSPHATE/PHOSPHATE TRANSLOCATOR 2	0.006368	21
AT4G31800	WRKY18	WRKY TRANSCRIPTION FACTOR 18	0.006472	144
AT2G47890	AT2G47890	ZINC FINGER PROTEIN CONSTANS-LIKE 13	0.006472	131
AT3G26520	TIP2	TONOPLAST INTRINSIC PROTEIN 2	0.006759	30
AT2G26980	CIPK3	CBL-INTERACTING PROTEIN KINASE 3	0.006759	19
AT2G29060	AT2G29060	SCARECROW-LIKE PROTEIN 34	0.006759	3
AT3G18100	MYB4R1	MYB DOMAIN PROTEIN 4R1	0.006954	0
AT2G45130	SPX3	SPX DOMAIN GENE 3	0.00766	0
AT4G16265	NRPE9B	RNA POLYMERASES M/15 KD SUBUNIT	0.007859	12
AT5G58620	AT5G58620	ZINC FINGER CCCH DOMAIN-CONTAINING PROTEIN 66	0.008259	12
AT2G20180	PIF1	PHY-INTERACTING FACTOR 1	0.008259	3
AT5G61900	BON	CALCIUM-DEPENDENT PHOSPHOLIPID-BINDING COPINE-LIKE PROTEIN	0.008809	66
AT3G50700	ATIDD2	ARABIDOPSIS THALIANA INDETERMINATE (ID)-DOMAIN 2	0.009543	7
AT1G06160	ORA59	OCTADECANOID-RESPONSIVE ARABIDOPSIS AP2/ERF 59	0.0099	8
AT4G14220	RHF1A	RING-H2 GROUP F1A	0.0099	32

Table 2.8: Regulatory genes (GO:0065007) predicted to be involved in the response to water deprivation (GO:0009414) based on the residuals co-differential expression network, at FDR = 0.01. The last column gives the number of sample networks that support the residuals prediction at FDR = 0.01. We screened literature for direct or indirect evidence supporting the top-10 predictions. Predictions supported indirectly (evidence for involvement in related processes or direct evidence for homologs in other species) are highlighted in yellow. Relevant references are indicated by superscripts in the second column.

ORF	Name	Description	Corrected P Value	# Sample Network Predictions
AT5G47220	ERF2	ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 2	1.72E-10	11
AT5G39610	ATNAC2	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 2	7.57E-09	115
AT1G09530	POC1	PHOTOCURRENT 1	1.95E-07	40
AT4G17500	ATERF-1	ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 1	1.32E-06	84
AT5G27520	PNC2	PEROXISOMAL ADENINE NUCLEOTIDE CARRIER 2	2.41E-05	109
AT5G13220	JAS1	JASMONATE-ASSOCIATED 1	4.57E-05	11
AT2G17040	ANAC036	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 36	8.52E-05	14
AT1G08450	CRT3 [106]	CALRETICULIN 3	8.77E-05	16
AT5G13820	TBP1	TELOMERIC DNA BINDING PROTEIN 1	1.23E-04	48
AT1G19180	JAZ1	JASMONATE-ZIM-DOMAIN PROTEIN 1	1.78E-04	157
AT2G30140	AT2G30140	UDP-GLUCORONOSYL/UDP-GLUCOSYL TRANSFERASE-LIKE PROTEIN	1.93E-04	118
AT3G29035	ATNAC3	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 3	4.92E-04	7
AT1G60190	PUB19	U-BOX DOMAIN-CONTAINING PROTEIN 19	5.91E-04	566
AT5G26920	CBP60G	CAM-BINDING PROTEIN 60-LIKE G	6.02E-04	16
AT3G02875	ILR1	IAA-LEUCINE RESISTANT 1	9.55E-04	31
AT2G47890	AT2G47890	ZINC FINGER PROTEIN CONSTANS-LIKE 13	0.001192	147
AT1G21410	SKP2A	F-BOX PROTEIN SKP2A	0.001721	148
AT5G22290	ANAC089	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 89	0.001721	169
AT5G13080	WRKY75	PUTATIVE WRKY TRANSCRIPTION FACTOR 75	0.002271	63
AT1G51140	AT1G51140	TRANSCRIPTION FACTOR BHLH122	0.002367	189
AT5G24470	PRR5	PSEUDO-RESPONSE REGULATOR 5	0.002896	58
AT4G26080	ABI1	ABA INSENSITIVE 1	0.002951	489
AT3G61190	BAP1	BON ASSOCIATION PROTEIN 1	0.002951	142
AT4G17230	SCL13	SCARECROW-LIKE 13	0.002951	161
AT2G19450	RDS1	DIACYLGLYCEROL O-ACYLTRANSFERASE 1	0.002951	35
AT1G51660	MKK4	MITOGEN-ACTIVATED PROTEIN KINASE KINASE 4	0.002951	16
AT3G59700	LECRK1	LECTIN-RECEPTOR KINASE 1	0.003236	23
AT3G55980	SZF1	SALT-INDUCIBLE ZINC FINGER 1	0.003641	206
AT3G01320	SNL1	SIN3-LIKE 1	0.003860	1
AT1G29400	AML5	ARABIDOPSIS MEI2-LIKE PROTEIN 5	0.004384	59
AT1G59580	MPK2	MITOGEN-ACTIVATED PROTEIN KINASE HOMOLOG 2	0.004426	6
AT2G34720	NF-YA4	NUCLEAR FACTOR Y, SUBUNIT A4	0.004521	48
AT1G19270	DA1	DA 1	0.005050	13
AT1G78600	LZF1	LIGHT-REGULATED ZINC FINGER PROTEIN 1	0.005283	58
AT1G74670	AT1G74670	PUTATIVE GIBBERELLIN-REGULATED PROTEIN	0.005584	5
AT3G50650	AT3G50650	SCARECROW-LIKE PROTEIN 7	0.005584	6
AT4G31800	WRKY18	WRKY TRANSCRIPTION FACTOR 18	0.005633	46
AT4G11260	EDM1	ENHANCED DOWNY MILDEW 1	0.006278	2
AT5G14960	E2L1	DP-E2F-LIKE 2	0.006278	14
AT1G75410	BLH3	BEL1-LIKE HOMEODOMAIN 3	0.006278	18
AT1G28370	ERF11	ERF DOMAIN PROTEIN 11	0.006821	174
AT4G34390	XLG2	EXTRA-LARGE GTP-BINDING PROTEIN 2	0.007503	26
AT1G01560	ATMPK11	MITOGEN-ACTIVATED PROTEIN KINASE 11	0.007658	21
AT3G08720	S6K2	ARABIDOPSIS THALIANA SERINE/THREONINE PROTEIN KINASE 2	0.007658	121
AT5G01540	LECRKA4.1	LECTIN-DOMAIN CONTAINING RECEPTOR KINASE A4.1	0.007948	25
AT1G56650	SIAA1	SUC-INDUCED ANTHOCYANIN ACCUMULATION 1	0.007951	10
AT4G36900	RAP2.10	RELATED TO AP2 10	0.008623	102
AT4G18890	BEH3	BES1/BZR1 HOMOLOG 3	0.008623	1
AT4G32010	VAL2	VP1/ABI3-LIKE 2	0.008958	0
AT5G67450	AZF1	ARABIDOPSIS ZINC-FINGER PROTEIN 1	0.009371	3
AT4G12720	GFG1	GROWTH FACTOR GENE 1	0.009526	28
AT5G58620	AT5G58620	ZINC FINGER CCCH DOMAIN-CONTAINING PROTEIN 66	0.009818	7

2.3 Discussion

We reanalysed a set of gene expression profiles of single wild-type *Arabidopsis* leaves of three accessions grown in tightly controlled growth room conditions across six labs. We focused on the residual expression differences that remain among the profiled leaves after controlling for lab and/or accession-dependent gene expression effects. Intriguingly, these residuals, generally considered experimental noise, still harbour a remarkable amount of biologically relevant expression variation, comparable to the information content of same-sized expression compendia incorporating traditional large-effect perturbations on pooled plant samples. Our analyses show that the expression variations among the individual plants are not random, but most likely reflect subtle differences in their growth environment or, due to slight differences in the developmental stage of sampled leaves or from population-level mechanisms to cope with stress[115], in spite of the detailed protocol used to control the experimental growth conditions[4]. In support of this notion, many of the stress responses to environmental factors that are difficult to rigorously homogenise in even the best of experimental setups, such as salt, water, and infestations by fungi, score above average in our gene function prediction performance assessment, while responses to factors that are more easily controlled or homogenised across plants in lab conditions, such as oxygen levels, light intensity, UV, and insects, score below average. In between these extremes are responses to factors that may have been controlled to an intermediate extent in the original setup, such as temperature, oxidative stress, mechanical stimulus (e.g., through plant handling), and starvation[4]. Responses to relatively harsh stresses, such as desiccation, which arguably did not impact the lab-grown plants in the original experiment[4], score comparatively worse than responses to milder or more generally defined stresses, such as water deprivation. In addition, processes that are thought to have a low impact on gene expression in fully expanded leaves as profiled in the original study[4] (e.g., cell cycle, cell differentiation, and auxin and brassinosteroid signaling) are generally not well represented in the gene network learned from the residuals data set, whereas several hormone signaling pathways associated with responses to various biotic and abiotic stresses (JA, ABA, and ethylene) score well above average.

In addition to assessing its capacity to recapitulate known gene functions, we used the residuals data set to predict the involvement of novel genes in regulating six of the best performing processes in our prediction performance screen, and we sought to experimentally validate the top predicted novel regulator of the JA signaling response, ILL6. We found increased phenotypic sensitivity to exogenous jasmonate, increased wound-induced JA-Ile accumulation in *ill6* mutants versus wild-type plants, and a decreased capacity to release Ile from exogenously applied JA-Ile, consistent with a negative regulatory role of ILL6 in the jasmonate response. These results highlight the role of jasmonate as a sentinel of environmental stress and, more generally, show that expression responses to uncontrolled subtle variations in plant growth conditions can be used effectively to point to novel regulatory relationships.

Noisy gene expression caused by variability in environmental parameters or intracellular stochastic effects is often considered a nuisance, although some authors have recently used intrinsic expression noise propagation to decipher regulatory influences in single-celled organisms[116–118]. It is currently impossible to assess which proportion of the residuals is due to true stochastic variation emanating from the stochastic nature of cellular processes, instead of micro-environmental variation, as the two are impossible to separate in the setup used by[4]. Even if it were possible to separate inherent stochastic

effects from micro-environmental effects, it is unclear to what extent inherent stochastic variations on the cellular level, if they would propagate through the cellular regulation network, would contribute to coordinated expression variation across genes in the context of a multicellular organism, as they would likely be averaged out to some degree across all cells in an individual plant or leaf. As outlined above, our results suggest that the observed residual expression variation derives mostly from subtle variations in the micro-environmental growth conditions of individual plants and that this expression "noise" contains valuable information on the wiring of biological networks, on par with the amount of information that can be extracted from controlled perturbations. In the prevailing perception of the scientific method, the stochastic features of uncontrolled experimental setups could be considered diametrically opposed to the experimental design features needed to ensure reproducibility. In the classical view, reproducibility is understood as the capacity to obtain the same results under the same controlled conditions. But from a systems biology perspective, reproducibility may be assessed on a different level. Reproducibility of a reverse-engineered gene network entails that the same interconnections among genes can be recovered from comparable data sets, which in this context are not necessarily copies that are systematically generated under exactly the same conditions. In fact, for large-scale gene network inference, the exact nature of the experimental conditions is secondary in importance to the requirement that similar conditions occur across the condition set when performing repeat experiments. In this respect, profiling the expression response of individuals to uncontrolled conditions can be regarded as sampling from a multivariate probability distribution, with each dimension being a random environmental factor. Given a large enough sample size, the effect size distributions in uncontrolled expression profiling experiments should therefore essentially be reproducible and so should the gene networks recovered from them.

The data set reanalysed here contained only a limited sample of 41 individuals, resulting in poor function prediction F-measures in the range 0 to 0.4. In addition, the data set was suboptimal because of the multiple ecotypes and labs involved in the original study[4], leading to systematic biases that may not have been pruned out entirely by ANOVA analysis. Nevertheless, it is clear that the uncontrolled residuals contain a significant amount of information on the underlying gene network structure. The results presented here suggest that expression profiling of wild-type individuals under uncontrolled conditions should be considered as an alternative data generation strategy for unraveling the wiring of biological networks. Algorithms used for this purpose are notoriously data-demanding, to the extent that unraveling a substantial part of an organism's transcriptional wiring easily requires hundreds of independent, controlled perturbations[119–123]. Given the substantial resource and time expenditure associated with controlling growth conditions and treatments, generating mutant lines, and profiling biological replicates, profiling uncontrolled individuals may prove more cost-effective for generating sufficient amounts of data for large-scale reverse engineering efforts.

In addition, uncontrolled data sets are fundamentally different from traditional data sets with respect to the perturbation structure across experimental conditions. In traditional data sets, only a single major perturbation is usually applied in any given experiment, while in an uncontrolled data set, multiple unidentified (mild) perturbations may impact the expression profile of an individual simultaneously. For instance, an individual plant may have been subjected to both watering and temperature conditions that are subtly different from its neighbours. This multifactorial setup is exactly the setup encountered by plants in the field, where they are irregularly and often simultaneously impacted by several abiotic and

biotic stresses, the responses to which often operate in synergistic or antagonistic interaction to modulate plant fitness. In this respect, uncontrolled field data sets screening multifactorial phenotypic responses under natural variation in the growth environment may prove useful to identify and quantify crosstalk between pathways, an issue that is not easily tackled in a lab environment but is of paramount importance for predicting the phenotypic effects of candidate yield or stress tolerance-enhancing mutations in the field. Although the use of natural variation on the genotype level has become mainstream in recent years, e.g. in genome-wide association studies and expression quantitative-trait-locus (eQTL) analyses[124–128], the potential use of natural variation in gene expression triggered by variations in environmental conditions has only recently begun to gather attention[129,130]. In most species, natural variation other than on the genotype level is still considered a nuisance rather than a potential asset. However, our results suggest that sampling natural environmental variation may be of general use for reverse engineering genetic networks, not only in plants, but also in species such as human, for which uncontrolled environmental variation is largely unavoidable and controlling experimental conditions and treatments is often impossible due to ethical constraints.

2.4 Methods

2.4.1 Data Sets and Extraction of Co-differential Expression Networks

Raw microarray data for 41 individual *Arabidopsis thaliana* leaves[4], profiled using the AGRONOMICS1 microarray platform[28], were obtained from the AGRON-OMICS repository (`http://www.agron-omics.eu/`). The raw data were RMA normalized using the Bioconductor R package, version 2.5[35]. We retained only the Affymetrix ATH1 probe sets present on the AGRONOMICS1 array for calculating gene expression levels (using the agronomics1_ath1probes.cdf file), to facilitate comparisons between this data set and the sampled data sets for pooled plants (see below). The log-transformed expression profiles were subjected to gene-specific ANOVA models of the form:

$$E_{ijk} = \mu + L_f + A_k + (LA)_{jk} + \epsilon_{ink} \tag{2.1}$$

with i (= 1..41) indexing the number of expression values obtained per gene, μ the baseline expression level of a given gene, L_j the lab effect (j = 1..6), A_k the accession effect (k = 1..3), LA_{jk} the lab $\times$ accession interaction, and ϵ_{ijk} the residual error on the log expression level. The residuals ϵ_{ijk} were used for all further analyses. Table 2.1 indicates the numbers of samples on which unbalanced design ANOVA estimation of lab, accession, and lab $\times$ accession effects was based. Although the overall number of data points is limited, the numbers of leaves are fairly balanced across labs and accessions, and with one exception, there are always three data points to estimate a particular interaction effect.

To construct same-sized sample data sets on perturbed and pooled plants, 688 Affymetrix ATH1 microarray experiments profiling the response to various perturbations on leaf and shoot tissues were extracted from the CORNET database[131], and the resulting compendium was randomly sampled without replacement to obtain 1000 data sets containing 41 experiments each. These were preprocessed as described above, and expression ratios (perturbations versus their respective control conditions) for

19,937 *Arabidopsis* genes were obtained using a custom cdf file designed to minimise cross-hybridisation effects[132]. In all data sets, only the 19,760 nuclear genes in common between the AGRONOMICS1 and ATH1 cdf files were retained for further analysis.

Co-differential expression networks and expression modules for the residuals data set and sample data sets were obtained using ENIGMA 1.1[1]. ENIGMA requires the definition of up- and down- regulation thresholds, either based on differential expression P values or expression log ratio thresholds. Since differential expression P values can by design only be computed for the sample data sets, but not for the residuals data set, we standardised the treatment of all data sets using a log ratio threshold of 0.3498 to define up- and down-regulation of gene expression (see Section 2.2). Note that the residuals can also be considered log ratios with respect to the baseline expression level of a gene over all leaves after correcting for lab and accession effects (Equation 2.1). The FDR level for detecting significant co-differential expression links was set to 0.01. For functional annotation on the level of expression modules, GO ontology information and annotations for *Arabidopsis* were obtained from the GO database (`www.geneontology.org`, annotation version 10/23/2012), and annotations with non-experimental evidence codes IEA, ISS, and RCA were discarded. GO enrichment of gene modules was assessed using hypergeometric tests, and the resulting P values were corrected for multiple testing using the Benjamini and Hochberg FDR correction at FDR = 0.05. Potential regulators of a module were predicted from the set of genes annotated to "biological regulation" in GO (GO:0065007) at FDR = 0.01. The remaining ENIGMA parameters were set to default values. For use in gene function predictions, negative correlation edges were removed from the co-differential expression networks. Basic network topology parameters (network density and clustering coefficient for the major connected component of each network) were obtained using NetworkX 2.6.4 (`http://networkx.github.com/`).

2.4.2 Gene Function Prediction

We predicted the function of a given gene from a given network by performing GO enrichment analysis on its network neighbourhood using a custom-tailored derivative of PiNGO, a software tool to screen biological networks for genes that may be involved in a process of interest[133]. Gene functions were predicted with hypergeometric tests, and the resulting P values were corrected (per network) with the Benjamini and Hochberg multiple testing correction. The resulting GO predictions were then compared with the known GO annotations, and precision, recall, and F-measure (harmonic mean of precision and recall) were scored for every network for a wide array of GO categories (Figure 2.6) at prediction FDR thresholds ranging from 10e-2 to 10e-11. For every functional category, the relative prediction performance of the residuals network with respect to the sample networks was classified as very good, good, average, poor, or very poor (see Figure 2.4 legend) based on the root mean square deviation of the residuals network F-measures from the 25th, 50th, and 75th percentiles of the sample network F-measures over the FDR subrange in which the residuals network exhibited defined F-measures, with deviations normalized to the square root of the residuals F-measure.

The global function prediction performance of a given network was calculated using a gene-centric method described by[93], based on assessing the overlap between predicted and annotated GO functional paths for a given gene (i.e., the path from an annotated or predicted GO term to the root of the GO

hierarchy), while taking into account the depth of predictions and annotations in the hierarchical GO structure. Recall and precision were calculated for every gene as described[93]. The overall prediction recall and precision score of an entire gene network are then defined as the arithmetic mean of the recall and precision values across all genes. Recall, precision, and F-measure were calculated for every network at prediction FDR thresholds ranging from 10e-2 to 10e-11.

2.4.3 JA Signaling Response Gene Prediction

PiNGO[133] was used to screen all networks for known regulators that are potentially involved in the JA signaling response. To obtain high-confidence functional predictions, computationally derived GO annotations with evidence codes IEA, ISS, and RCA were discarded. The set of 19,760 genes present in all data sets was used as the reference set. "Biological regulation" (GO:0065007) was set as the "start" GO category, while "response to JA stimulus" (GO:0009753) was used as the "target" and "filter" GO category. P values were calculated with hypergeometric tests and corrected with the Benjamini and Hochberg multiple testing correction at FDR = 0.01. The same protocol was used for predicting novel regulators for the other processes listed in Tables 2.4 to 2.8, with the "target" and "filter" GO categories defined accordingly.

2.4.4 Plant Material, Growth Conditions, and Genetic Analysis

Plants were grown at 22 °C in Sunshine Mix LC1 potting soil (wounding experiments) or Jiffy 7 peat pellets (in vivo hydrolysis assays; Jiffy Products) and 10 h (wounding, in vivo hydrolysis assays) or 16 h (growth inhibition assay) of light at 100 to 120 μmol/photons/m^2/s. *Arabidopsis* accession Col-0 was obtained from the ABRC (ABRC stock CS70000). The *ill6* mutant lines were derived from ABRC stocks Salk_024894C (*ill6-1*) and CS852193 (*ill6-2*), both in the Col-0 background. To identify homozygous T-DNA insertion mutants, genomic DNA of individual plants of these lines was used as template in a three-primer PCR reaction. *ILL6* transcript accumulation in these lines was examined by RT-PCR. The sequences of primers used in these analyses are included in Table 2.9.

Table 2.9: Sequences of oligonucleotide primers used for *ILL6* PCR analyses.

Primer	Sequence (5'-3')
P1	TTA TGA ATG TTT ATC ATT TAA GTA TCT CTC AGC CAC GGC
P2	CGC ACC TCT TGA ATA CGT TTC
P3	GAC TAT GCT TCT TGG TGC TGC
P4	CACC ATG GAC AAT CTC CGG AAA CTT AAT CTT CTC TCT G
P5 (pROK2 LB)	TGG AAC AAC ACT CAA CCC TAT CTC GG
P6 (pDsLox LB)	AAC GTC CGC AAT GTG TTA TTA AGT TGT C
ACT2-f	CTG GAT TCT GGT GAT GGT GTG TC
ACT2-r	TCT TTG CTC ATA CGG TCA GCG

2.4.5 Growth Inhibition Assay

Surface-sterilised and cold-stratified seeds were plated on half-strength Murashige and Skoog media, pH 5.8, containing 0.8% Suc, 0.8% agar, and 0.5 g/L MES. After 3 to 4 d, seedlings of equal root length ($\sim$ 1

cm) were transferred to plates of the same media containing various concentrations of MeJA or an equal volume of carrier (DMSO); to reduce inter-assay variability, these plates were always allowed to air-dry in a laminar flow hood for exactly 1 h. Each plate contained an equal number of seedlings of all three genotypes. In the data presented in Figures 2.9-A and 2.9-B, the seedlings were transferred to three replicate plates per concentration of MeJA and each replicate was placed on a separate shelf of a plant growth chamber. After 8 d on JA-containing media, the length of the primary root of each seedling was measured, and the shoot tissue was removed and weighed. A minimum of 16 seedlings was analysed for each genotype at each concentration.

A linear mixed model was fitted to the data and analysed using the residual maximum likelihood method. The model included fixed effects due to genotype, MeJA concentration, their interaction, and random effects due to the replicate plate and shelf. The significance of fixed effects was judged by *F*-test. Differential sensitivity of the mutant's root elongation and shoot weight were seen in other independent experiments.

2.4.6 Wounding Treatments and JA-Ile Analysis

Thirty-day-old plants of Col-0, *ill6-1*, and *ill6-2* were wounded evenly with a hemostat twice across the width of each of three fully expanded leaves, crushing 40 to 50% of the leaf surface area. At various time points after wounding, 200 to 300 mg of damaged leaves from two individual plants was harvested together and immediately frozen in liquid nitrogen and stored at 280 °C until jasmonate extraction.

Extraction and quantification of endogenous JA-Ile from plant tissue were according to previously described methods[134,135]. A known amount of [$^{13}C_6$]JA-Ile was added to the frozen samples at the beginning of extraction as an internal standard. Compounds were separated on an Ascentis C18 column (1.7 μM, 2.1 × 3 × 50 mm) using an Acquity ultraperformance liquid chromatography system (Waters). A Quattro Premier XE tandem quadrupole mass spectrometer (Waters) was used in an electrospray negative mode to detect JA-Ile (322 → 130) and [$^{13}C_6$]JA-Ile (328 → 136).

The data from two independent experiments were analysed together. A linear mixed model was fitted to the data and analysed using residual maximum likelihood, including fixed effects due to genotype, time, their interaction, and random effects due to the replicated experiments. The significance of fixed effects was judged by *F*-test.

2.4.7 JA-[^{14}C]Ile Synthesis and in Vivo Hydrolysis Assay

JA was obtained by base-catalysed hydrolysis[136] of MeJA (Bedoukian Research) and purified by reverse-phase HPLC[137]. For synthesis of JA-[^{14}C]Ile, JA (14 mg), $_L$-Ile (8 mg), and $_L$-[^{14}C]Ile (5.5 μCi, specific activity 55 mCi/mmol; American Radiolabeled Chemicals) were coupled and purified by open-column silica chromatography as detailed[138]. For plant treatments, 50,000 dpm of JA-[^{14}C] Ile in an aqueous 20% DMSO solution was applied in a single 10-mL drop to individual leaves of individual plants. After 24 h, leaves were excised and extracted individually in 4 mL of 95% ethanol at 70 °C for 45 min. These extracts were dried under a stream of nitrogen, resuspended in 50 mL of 95% ethanol, and separated by thin layer chromatography (silica gel 60; EMD Millipore) in chloroform:methanol:acetic acid (70:30:2, v:v:v). Radioactivity was detected with a Typhoon FLA 7000 phosphor imager (GE Healthcare Life

Sciences). Images were background subtracted and bands quantified using ImageJ[139]. [^{14}C]Ile was identified by cochromatography with an authentic standard. The $\log_{10}$- transformed data of Figure 2.9-E were analysed by one-way ANOVA, and the significance of the genotype effect was judged by *F*-test. The $\log_{10}$- transformed data of Figure 2.9-F were analysed by Student's *t* test.

2.5 Accession Number

Sequence data from this article can be found in the *Arabidopsis* Genome Initiative or GenBank/EMBL data libraries under accession number At1g44350/NM_103546.3 (*ILL6*).

2.6 Acknowledgements

We thank two anonymous reviewers for insightful comments on the article. This research was supported in part by Fund for Scientific Research-Flanders Grant G.0029.11 to S.M., National Institutes of Health Grant R01 GM57795 to G.A.H., U.S. Department of Energy Grant DE-FG02-99ER20323 to J.B., and by the Integrated Project AGRON-OMICS, in the Sixth Framework Program of the European Commission (LSHG-CT-2006-037704). S.M. is a fellow of the Fund for Scientific Research-Flanders.

2.7 Author contributions

S.M. conceived the study. R.B., J.H. and S.M. designed the research. R.B. and S.M. performed computational analyses. J.H., T.M. and S.M. supervised computational analyses. M.V. performed statistical Micro-Environmental Expression Variation analyses. R.B., P.H., A.G and S.M. interpreted data. J.B. and G.A.H. designed and supervised the JA signaling experiments. J.B.J. and A.J.K.K. performed and analysed the JA signaling experiments. R.B. and S.M. wrote the manuscript and remaining authors contributed to reviewing and improving the article.

Chapter 3

Developmental route map of the endocycle in *Arabidopsis thaliana*

Rahul Bhosale, Steven Maere and Lieven De Veylder (*In preparation*).

Summary

Endocycle is a variant of the mitotic cell cycle during which cells repeatedly replicate their genome without going into mitosis, resulting into cellular polyploidy. In plant model organism *Arabidopsis*, polyploid cells are often seen in various developing organs such as leaves, trichomes, root, hypocotyl, etc. throughout its life cycle. In the past decade, researchers have studied these organs individually and identified many genes involved in the onset and progression of the endocycle. Yet, we know little about differential regulation of the endocycle in these organs during their development. However, new accumulating evidences are providing insights into the molecular control of developmental and environmental cues on the endocycle regulation and its role in the maintenance of apt growth rate and architecture of the organ.

For the author contributions, see page 64.

3.1 Cell cycle and endocycle

Organism development involves the continuous and reiterative organogenesis during which complex developmental programs maintain the production of new cells and subsequent differentiation. For instance in plants, shoot and root apical meristems are the sites where new cells actively proliferate through the mitotic cell cycle. Upon leaving the meristem, cells start to differentiate and simultaneously increase their cell size through post-mitotic expansion (Figure 3.1-a). The switch from proliferation to differentiation at the meristems is accompanied by the transitions from the mitotic cell cycle to the endocycle, an alternative cell cycle during which cells duplicate their genome without cell division[140,141]. Every round of genome doubling increases the nuclear content (i.e. endoploidy) of the cells (Figure 3.1-b), e.g. in plant model organism *Arabidopsis*, many cell types in leaves, roots and hypocotyls reach 16C or 32C (C = haploid DNA content)[142]. For instance, the flow-cytometry endoploidy profile of *Arabidopsis* root is represented in Figure 3.1-c. The recent progress in understanding of the molecular machinery in mitotic cell cycle and endocycle suggest that both these cell cycles share many key components.

3.1.1 Mitotic cell cycle machinery

The mitotic cell cycle involves a rapid sequence of DNA synthesis phase (S-phase) and mitosis (M-phase), which are preceded by two gap phases G1 and G2 respectively. Cell cycle progression is controlled by cyclin-dependent kinases (CDKs) and their interacting partners, cyclins (CYCs), which regulate the kinase activity of CDKs[141,143]. Plant have two main classes of CDKs that directly regulate the cell cycle, (i) CDKA, which has functional homolog in yeast Cdc2/Cdc28p, and (ii) CDKB, which has been found in only plants. In *Arabidopsis*, CDKA is encoded by a single gene named *CDKA;1*, whereas there are two homologs for each *CDKB*, designated as *CDKB1;1*, *CDKB1;2*, *CDKB2;1*, and *CDKB2;2*. Recent studies suggest role of CDKA for both the G1-S and the G2-M transition, while CDKB1 and CDKB2 appears to function for the S-G2-M transition and G2-M transition, respectively[140,144]. There are at least 32 CYCs [10 A-type (CYCA), 11 B-type(CYCB) , 10 D-type(CYCD) and 1 H-type(CYCH)] with a putative role in the cell cycle progression. CYCD are known to regulate the G1-S transition, whereas CYCA regulate S-M phase control and CYCB both G2-M transitions and intra-M-phase control[140,141,143].

The expression of CYCD often depends on the stimulation by plant hormones, growth conditions and development[140]. These mitogens trigger the production of CYCD which in turn activates CDKs during late G1 phase and push cells into the next phase of cell cycle i.e. G1-S transition point, which is controlled by Adenovirus E2 promoter binding factor (E2F)/retinoblastoma-related (RBR) pathway. These CDK's phosphorylate the RB protein at multiple sites resulting in the inactivation of RB and the release of active E2F-Dimerisation Partner (DP) transcription factors that transcriptionally activates hundreds of E2F target genes, which are mostly DNA replication genes[145,146]. The E2F-DP/RBR pathway is highly conserved among higher eukaryotes. *Arabidopsis* genome encodes a total of six E2F factors, which can be subdivided (based on their structure and functional properties) into typical (E2Fa, E2Fb and E2Fc) and atypical (DP-E2F-LIKE1[DEL1]/E2Fe, DEL2/E2Fd, DEL3/E2Ff)[147]. Typical E2F factors need to dimerise with DP to gain a high DNA-binding specificity while atypical ones can bind as a monomers, as they possess two DNA-binding domains. Among typical E2Fs, both E2Fa and E2Fb are transcriptional

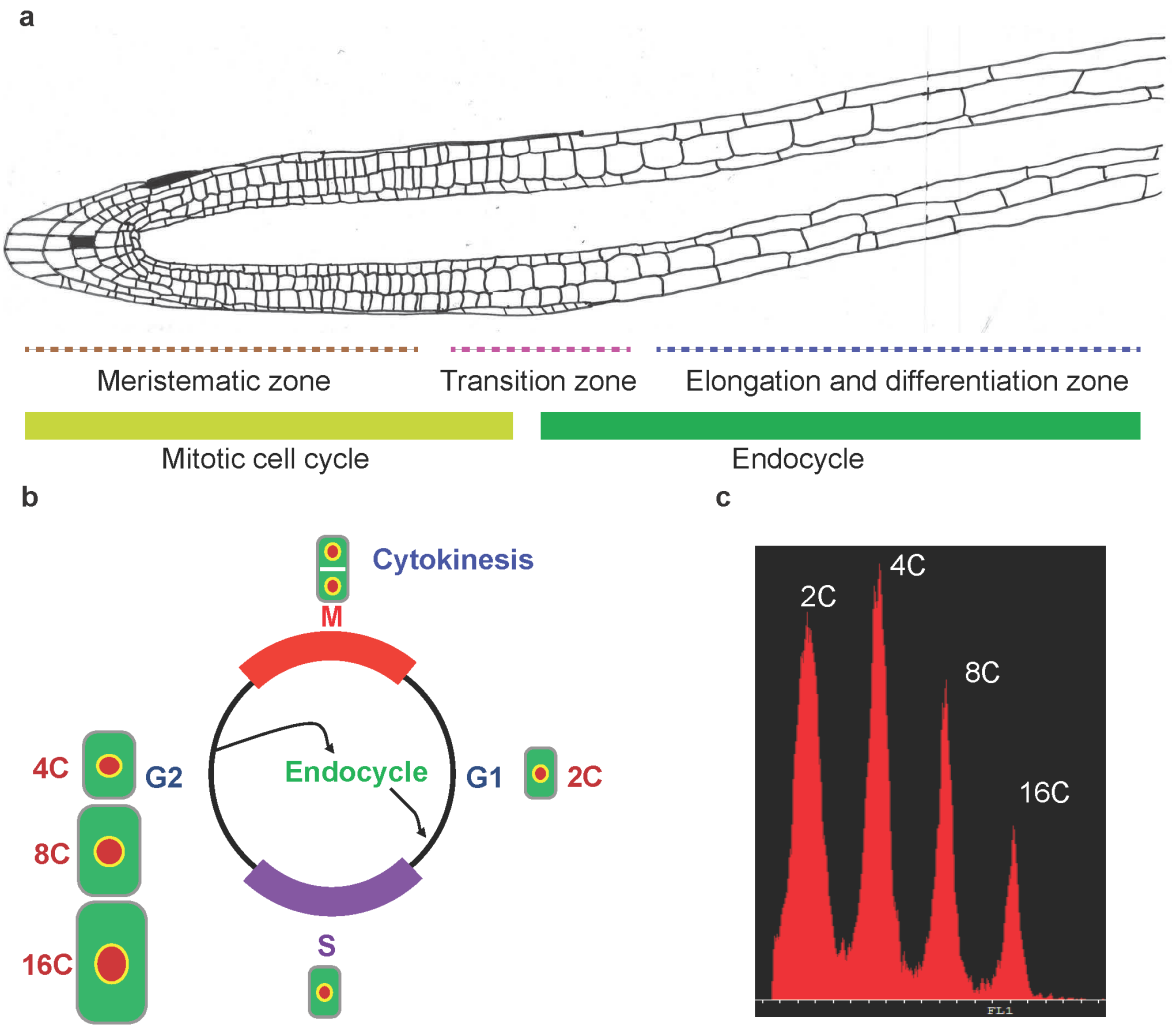


Figure 3.1: Representation of the cell cycle and endocycle zones in *Arabidopsis thaliana* root. (**a.**) During *Arabidopsis* root development, proliferating cells (by mitotic division) can be observed in meristematic zone, cells in mitotic to endocycle transition can be observed in elongation zone and endocycling cells can be observed in elongation and differentiation zone. (**b.**) During the classical mitotic cell cycle, DNA that is duplicated during the S phase is equally divided during the M phase, so that each daughter cell produced after cytokinesis possesses a genomic DNA content that is equal to that of its parents, being 2C (C equals the haploid DNA content). In this cycle, the S phase and the M phase are separated from each other by two intervening gap phases, G1 and G2. In contrast to the traditional mitotic cell cycle, during the endoreplication cycle (endocycle), no cytokinesis occurs between successive rounds of DNA replication. In this manner, the DNA content of the cell is doubled with every new round of DNA replication, resulting in the formation of cells with a DNA endoploidy level of 2C, 4C, 8C, 16C, 32C, etc. (**c.**) Flow cytometric endoploidy profile of *Arabidopsis* root shows population of cells at endoploidy levels 2C, 4C, 8C and 16C.

activators of the S-phase initiation and progression and potentially targets genes involved in DNA repair and chromatin dynamics, such as *CDC6*, *MCM3*, *ORC1*, *CDT1a*, *PCNA*, *RBR* and *RNR*[145,146]. In addition, they potentially target *CDKB1;1* gene that is required for entry into mitosis[148]. Contrary, E2Fc is a repressor with shortened C-terminal transactivation domain and thought to co-ordinate the cell cycle exit and cell division by working antagonistically to E2Fb. All atypical E2Fs (DEL1, DEL2 and DEL3) are considered as repressors because they lack a transcriptional activation domain[149,150].

The progression through G2-M is controlled by M-phase-specific activator (MSA) cis-acting element[151]. This MSA element is recognised by three Myb repeats (MYB3R) transcription factors to drive the expression of G2-M-phase specific genes. In *Arabidopsis*, there are five MYB3R proteins, among them MYB3R1 and MYB2R2 are known to positively control the expression of genes containing MSA element in their promoters such as *CDC20.1*, A-type (*CYCA1;1*) and B-type (*CYCB2;1*, *CYCB1;4*, *CYCB1;2*) CYC

genes and especially *KNOLLE* gene involved in cell plate formation during cytokinesis. The activity of MYB3R depends on their phosphorylation by CDK-CYC complexes and regulated in a feed forward loop manner, in which cyclins induced by the MYB3R proteins form a complex with CDKs that super-activate the MYB3R activity. Other MYB proteins such as CDC5 and MYB11 are also known to play a role in cell cycle progression but their mode of action is still unclear[152,153].

3.1.2 Switch to endocycle

The switch from mitotic cell cycle to the endocycle is assumed to be mediated by a reduction in M-phase-specific CDK activity. This CDK activity is controlled by the abundance of CYCs and their co-factors by at least three known mechanisms, (a) transcriptional regulation, where (i) CYCA2 are negatively regulated by a transcriptional repressor protein Increased Level of Polyploidy1-1D (ILP1) that leads to inactivation of CDKB1 G2-M specific activity[148] and (ii) G2-M expressed genes including CYCB are negatively regulated through lack of MYB3R phosphorylation due to repressed pre-mitotic CDK activity[154,155], (b) protein degradation, where G2-M specific regulators such as cyclins CYCB1;1 and CYCB1;2[156], as well as CYCA2;3 and CYCA3;1[157,158] are selectively marked for destruction through Cell Cycle Switch Protein 52 A2 (CCS52A)-anaphase-promoting complex/cyclosome (APC/C) ubiquitination that targets proteins to the 26S proteasome[159] and (c) Siamese-Related (SMR) family of plant specific CDK Inhibitor (CKIs)[160,161], possibly by inhibiting CDKA-CYCD3 complexes[162,163], CDKB1;1-CYC complexes or transcriptional repression of G2-M-phase genes by inhibiting MYB3R phosphorylation[161,162]. The progression of endocycle is regulated by the oscillating cycles of CDKA activity, which is required for DNA replication[164,165]. Recent studies suggest that the CDKA activity is regulated by Kip-Related Protein (KRP) family, a class of CKIs, which specifically binds CDKA-CYCD complexes in a dosage-dependent manner[166,167]. A low levels of KRPs inhibit the mitotic cell cycle, whereas a high levels arrest both the cell cycle and the endocycle.

3.1.3 Occurrences of the endocycle

In nature, endoreplication has been observed in a wide variety of cell types from lower invertebrates, arthropods, mammals to higher plants during various biological processes such as differentiation, growth, cell fate maintenance, metabolic activities, etc. Thus, the endocycle is recurrently proposed to be relevant in development of organisms. Endocycle occurrences and functions in these organisms are represented in Figure 3.2.

In lower invertebrates, endoreplication is often associated with increased cell and/or body size. In nematodes such as *Caenorhabditis elegans*, endoreplication occurs in syncytial hypodermal nuclei, and it is believed to be a crucial determinant of adult body size and body size evolution[168]. In the simple chordate *Oikopleura docoida*, striking spatial patterns of endopolyploidy have been documented. After early development, they undergo differential endocycling, which gives rise to endoploidy levels ranging from 4C to 1,300C and grows tenfold in mass[169]. In snails, terrestrial slugs and sea hares, giant neurons, epidermal gland cells and digestive gland cells undergo massive endoreplication and concordantly increase in size. These giant neurons can reach endoploidy levels of 260,000C. In arthropods such as crustaceans and insects, endoreplication is common, but extensively characterised in the fruit-fly

Drosophila melanogaster. In Drosophila, the endocycle is developmentally programmed, correlates with growth and increased metabolic activity and endoploidy levels range from 8C to 2,000C in different cell types such salivary glands, fat body epidermis, gut, trachea and renal tubules of fly larvae, and in neurons, glia, sensory bristles, gut and ovarian nurse and follicle cells of adult fly[170–173]. Endoreplication is also observed in the silk and poison glands of spiders such as *Pholcus phalangioides* and associated with their high protein output[174].

Endoreplication has been noted in mice[175] and humans, and probably occurs in most mammals. For instance, mammalian cells that endoreplicate include placental trophoblast giant cells (TGCs, upto 512C)[176], hepatocytes (upto 16C)[177], cardiac myocytes (4C-8C)[178], blood megakaryocytes[179] (upto 128C), epithelial keratinocytes[180], vascular smooth muscle cells[181] and primitive podocytes of the kidney[182]. Some of these cells endoreplicate during injury- or infection-mediated stress, whereas others endoreplicate as part of a developmental programme.

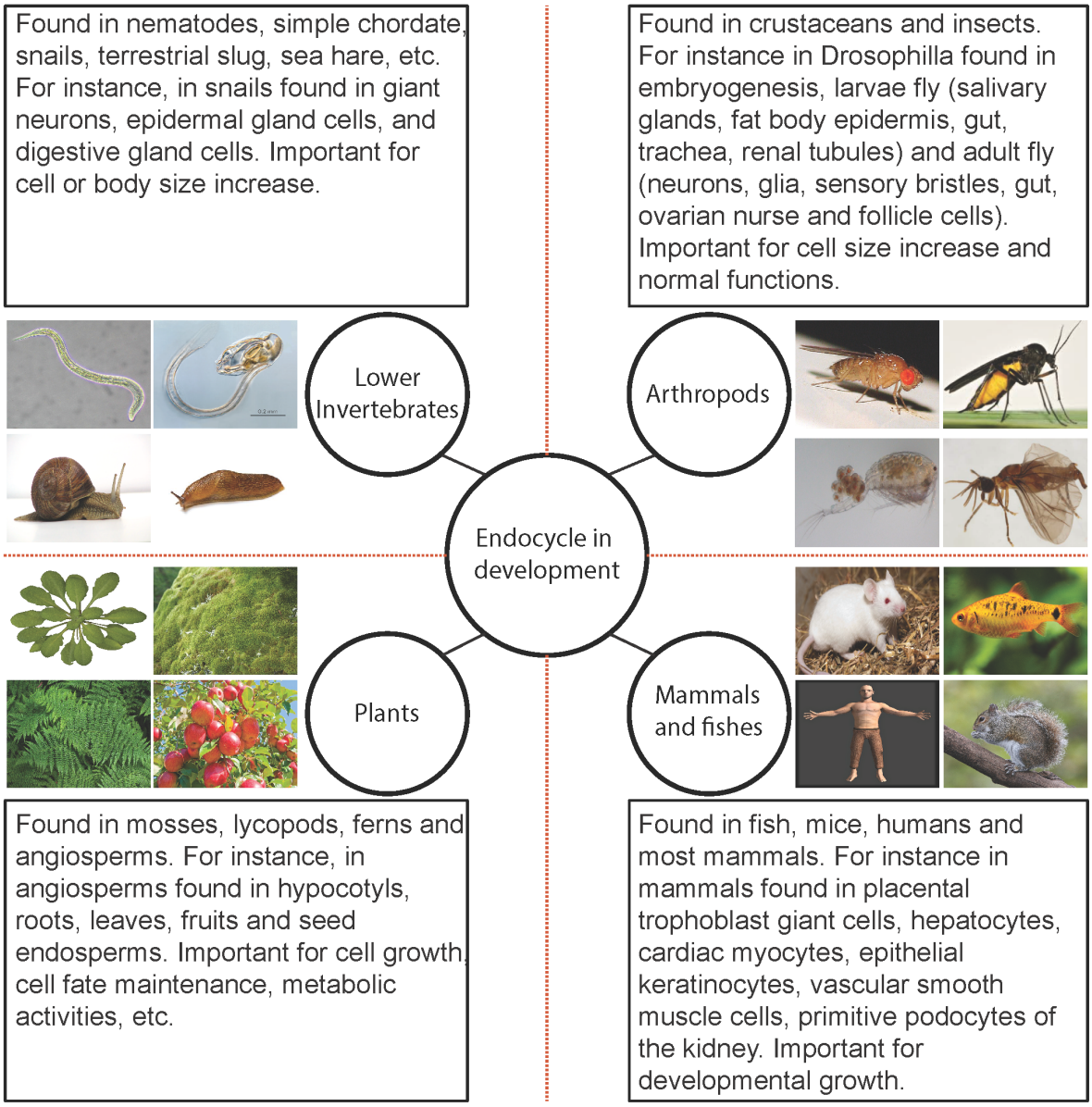


Figure 3.2: Schematic representation of endocycle occurrence during development of organisms in nature. In nature, endocycle is observed in a wide variety of cell types from lower invertebrates, arthropods, mammals to higher plants during various developmental stages of organisms. This figure was produced from the content primarily described in[183]

In plants such as mosses, lycopods, ferns and angiosperms, a wide variety of cell types undergo endocycle as an essential aspect of normal differentiation. In recent years, the plant model organism *Arabidopsis* has been extensively studied to understand the developmental and environmental control of endoreplication levels and their significance in plant growth and development.

3.2 Developmental control of the endocycle

In *Arabidopsis thaliana*, like in other higher eukaryotes, most of the structures and organs are formed post-embryonically through specialised regions called meristems. At the centre of these meristems are the pluripotent stem cells, which undergo cell division to maintain the proliferation rate with the cells undergoing differentiation upon leaving the meristem. The transition from cell proliferation to differentiation is accompanied by the transition from mitotic cell cycle to the endocycle. The balance between cell cycle and endocycle is co-ordinated tightly by developmental cues, which determines both the growth rate and architecture of the organ. In *Arabidopsis*, the endocycle is very common and polyploid cells are often seen during development of organs throughout its life cycle. For instance, the early growth of the plant after germination occurs in the soil where the hypocotyl grows primarily by endoreplicating cells and emerges out of the soil with two cotyledons. Further, the rosette leaves are formed and they progressively develop along with the special structures undergoing endoreplication to differentiate into trichomes. In later stages, the developing root quickly establishes three regions, the meristematic zone of mitotically dividing cells and the elongation and differentiation zone of endoreplicating cells (Figure 3.1). The developmental control on endoreplication during development of these organs is detailed below.

3.2.1 Hypocotyl development

The hypocotyl, a post-embryonic stem that connects two cotyledons and radicle, is responsible for the early growth of the seedling in soil. The hypocotyl development is characterised by the spatial and temporal regulation of the endoreplication. It involves only a few cell divisions after gemination and subsequent multiple rounds of endoreplication[184]3.3. On the first two days after germination, large portions of cells undergo up to 2-3 rounds of endoreplication. Between day 3-5, the hypocotyl undergoes one more round of endoreplication followed by exponential growth phase. Endoreplication is often associated with cell growth as DNA content analysis per cell suggests a correlation between cell size and endoreplication rates[185–187]. So, the additional round of endoreplication in dark is speculated to aid hypocotyl elongation in its search for light but still there is no clear evidence coupling these two facts. In addition, no drastic effects were observed on endoploidy levels of short hypocotyl mutant phenotypes[184,188]. The DNA content in cortex and epidermis is observed to be more than the stele and endodermis which indicates that the endoreplication during hypocotyl development is probably limited to these outer tissue-types. The cells in the central cylinder remains dividing as they participate in the thickening of the hypocotyl i.e., secondary growth[189]. Upon emergence from soil, endoreplication is negatively regulated by the sunlight (Figure 3.3). In recent years, *Arabidopsis* hypocotyls have been used as a model system to understand how the endocycle is regulated in dark and light conditions and whether the elongation growth is dependent or independent of endoploidy.

Physiological studies in dark/light treatment identified that light controls both hypocotyl elongation and endoreplication (see Section 3.3). Mutant studies are providing insights into how endocycle is regulated during development of hypocotyl in dark and light conditions. In light, the endoreplication is suppressed through the action of phytochromes (phys) and cryptochromes (crys), specifically phyA plays role in far-red light, phyB in red and white light while crys in blue light[184]. The phyA, phyB and elongated hypocotyl 4 (hy4, blue-light photoreceptor) mutants showed increased polyploid cells when grown in respective source of light[184]. In addition, COP1[184] and IPD1[188] has been shown to act downstream of the phytochromes and cryptochromes. Contrary, in dark COP1 acts by ubiquitylating the positive regulators of photomorphogenesis, such as the transcription factors HY5, Long Hypocotyl in Far-Red1 (HFR1) and Long After Far-red Light1 (LAF1) thereby targeting them for degradation by the proteasome. Recently, COP1 has been shown to interact physically and genetically with the MIDGET (MID) protein, a component of the Topoisomerase VI (TOPOVI) complex [composed of Root Hairless 2 (RHL2), *Arabidopsis thaliana* TOP6 Subunit B (AtTOP6B), MID and RHL1] that is necessary for the endoreplication in *Arabidopsis* providing the functional link between endoreplication and photomorphogenesis[190]. In another way, COP1 is also known to destabilise the E2Fb protein levels in dark which allows E2Fc to be present in abundance which in turn binds to *DEL1* promoter to reduce its transcription and to commence endoreplication[147].

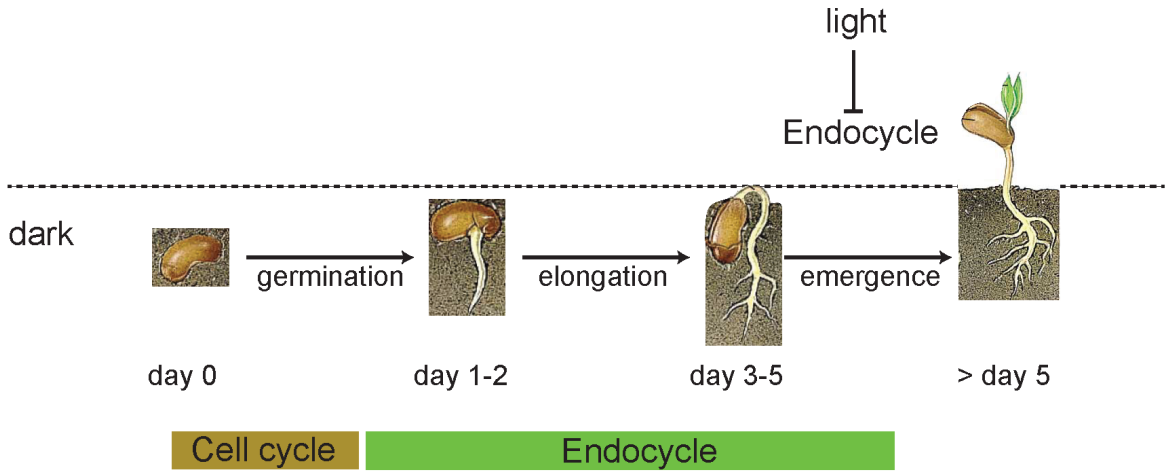


Figure 3.3: Endocycle during development of *Arabidopsis* hypocotyl. Hypocotyl growth in dark after seed germination (day 0) involves few cell divisions and multiple rounds of endoreplication. Upon emergence from the soil, endoreplication is negatively regulated by the sunlight. Two days post-germination, hypocotyl cells undergo 2-3 rounds of endoreplication and between 3-5 days these cells undergo an additional round of endoreplication. This figure is adapted from[191].

3.2.2 Leaf trichome development

Leaves are an important part of the plant as they play a pivotal role in photosynthesis, respiration and photo-perception. Their growth involves three distinct phases: leaf primordia development, primary and secondary morphogenesis. During primary morphogenesis, growth is sustained by successive cell divisions, and subsequently by cell expansion in secondary morphogenesis[192]. The transition from cell proliferation to expansion is often marked by the endocycle onset. Earlier, it was thought that the transition from cell proliferation to expansion proceeds in a gradient down the leaf, with cell proliferation

first ceasing in the tip and then progressively down the longitudinal axis[192]. Recently, the detailed kinematic and transcriptome analysis illustrated that it occurs abruptly and simultaneously with the onset of photomorphogenesis[193]. This study identified that *SMR1* was the only gene activated over the time course. SMR1 is highly similar to SIM, which has been shown previously to promote endoreplication. Later in leaf development, cells differentiate into distinct cell types such as guard cells, vascular tissue cells and trichomes, allowing them to perform various specialised functions[194]. Among these specialised cells, trichomes have been extensively studied for the endoreplication process. They are differentiated epidermal cells found on the aerial surfaces of nearly all plants. In various species, they adopt various morphologies and play a wide variety of functions including resisting insect herbivores, reducing transpiration, increasing freezing tolerance, protecting plants from UV light and aiding in precultivation seed dispersal[195].

In *Arabidopsis*, trichomes are unicellular, branched and non-glandular structures. Trichomes are first developed near the distal end of the maturing leaf and later proceeds basipetally. Trichomes are regularly spaced and has been found that their patterning relies on substrate-depletion and lateral inhibition mechanism[196]. Once the certain expression threshold of an activator complex [Glabrous1 (GL1), R2R3 MYB, GL3, TTG1, WD-40] in a cell is reached, the trichome fate is established and incipient trichome cell starts to express downstream genes such as *GL2* to regulate further outgrowth of the formation of typically three to four branches[197–199]. During this outgrowth trichome cell undergoes 3-4 rounds of endoreplication cycles, yielding a endoploidy level of 16C to 32C. The overview of the endocycle during trichome development is represented in the Figure 3.4. The DNA endoploidy levels and size of trichome cell are apparently correlated. Hence, they have been extensively studied to understand the role of endoreplication in cell growth and cell size. The mutants that have reduced endoreplication levels also show smaller trichomes with fewer branches while mutants with increased endoreplication levels have larger trichomes with more branches[200,201]. However, some experiments have shown contrasting results. The mutants with increased endoreplication showed no change in trichome cell size, whereas trichomes with enlarged cell size showed no change in nuclear content[202].

Nevertheless, recently it has been found that the endoploidy dependent trichome final cell size is controlled developmentally by the *GTL1*, GT-2 family trihelix transcription factor[203]. *GTL1* is not expressed at an early stages of trichome development when cells are still undergoing branching. Its transcription starts while cells reach their maximum size and directly binds to promoter *CCS52A1* and repress its expression to terminate the endocycle and hence the cell growth. The *gtl1* mutant shows that the trichomes are larger than those in the wild-type but they do stop growing, hence additional endoploidy-independent mechanisms are suggested to repress further cell growth[203]. In addition, the excess branching and endoreplication in *Arabidopsis* is controlled through the ubiquitination of one or more activators by *UPL3*, a member of HECT domain containing E3s[204]. *Arabidopsis upl3* mutation results in trichomes containing five or more branches instead of three and an additional round of endoreplication resulting in enlarged nuclei with endoploidy levels of up to 64C.

Other than the cell growth regulation, a recent study also identified that the endoreplication is crucial for the trichome cell fate maintenance[206]. The manipulation of endoreplication levels in trichome identified that the reduced levels of endoreplication results in reduced trichome numbers and identity loss. The dedifferentiating trichomes re-entered mitosis and re-integrated into the epidermal pavement-cell layer. Conversely, the promoted endoreplication using *CCS52A1* gene in glabrous patterning mutants

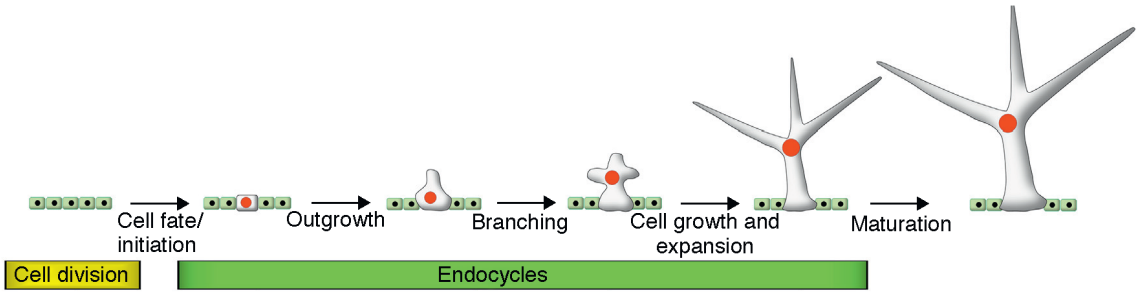


Figure 3.4: Endocycle during development of *Arabidopsis* trichome. In trichomes, mitotic-to-endocycle transition is developmentally regulated by the expression threshold of an activator complex. Once the trichome fate is established, downstream genes are expressed which in turn regulate the further outgrowth to form 3-4 branches. During this outgrowth, trichome cell undergo three to four rounds of endoreplication cycles, which correlates with the cell growth and expansion. Once the trichome cells matures and reach its maximum size, endocycle is terminated and further cell growth is repressed. This figure is adapted from [205].

has been shown to restore the trichome fate. CCS52A1 is known to co-operates with SIM to establish the endoreplication in *Arabidopsis* trichomes [156]. Besides, the trichome patterning genes such as *GL3* and *GL1* appears to have control on endoreplication. Loss of *GL3* function results in the reduction of endoreplication levels and conversely *try* mutants undergo one additional rounds of endoreplication. The role of GL1 might be debated as one study showed that over-expression in *GL1* results in increased endoreplication while other reports no effect. Recently, it has been found that SIM is a direct early target of GL3 and GL1. These regulators (GL3 and TRY) of trichome development are also shown to be dependent on the function of *CPR5* for their effects on trichome expansion and endoreplication [207].

So far, the trichome endoreplication cycle has been shown to be regulated by two core cell-cycle regulators. The first one is CDKA;1, the major regulator of mitotic cycles [164]. It is assumed that each round of endoreplication requires a cycle of alternating low and moderate CDK activity that ensures the licensing and activation of the replication origins, respectively [208]. CDKA activity remains unchanged with the onset of endoreplication and is essential for DNA replication. *Arabidopsis* weak loss-function *cdka;1* mutants exhibited smaller nuclei in trichomes along with reduction of trichome size and increase in cells in G1 and a reduction in endoreplication. The other cell cycle regulators involved in trichome cell-cycle control are CDK inhibitor SIM and KRP family members. The *Arabidopsis sim* mutant shows that endoreplication is partially converted into a mitotic program resulting in multicellular trichomes [160,161]. SIM likely targets CDK-CYCD complex, a misexpression of CYCD3;1 in trichomes has resulted in formation of multicellular trichomes [162]. KRP is known to bind CDKA-CYCD complexes. A weak expression of KRPs converted *Arabidopsis* multicellular mutant trichomes into endoreplicating single celled hairs, while lines that strongly overexpressed KRPs blocked both division and endoreplication [209].

3.2.3 Root development

Roots provide structural support to the aerial portions, acquire nutrients and water essential for plant growth, synthesise hormones, and are the site of interaction with soil bacteria [210]. Thus, the overall plant survival depends on appropriate root development, growth and function. The *Arabidopsis* root growth is an uniform and a continuous process. After the seed germination, the root apical meristem is established that produces new cells for the developing root. It contains a set of initial cells (stem

cells) that surround the quiescent centre (QC) that contains less mitotically active cells resting in an extended G1 phase. These stem cells give rise to different cell types confined in cell-files, which progress through three distinct developmental phases along the longitudinal axis on their way to maturity. In the meristematic zone, they divide multiple times to generate a pool of cells that consequently expand in elongation zone and differentiate in differentiation zone to acquire their specialised characteristics and functions[211]. The differentiation of cells is generally characterised by the onset of endoreplication as cells reach the meristematic and elongation zone border, they exit mitotic cell cycle and transits into the endocycle. In the past decade, physiological and mutant studies have speculated several roles and identified several molecular components of endoreplication involved in root development (see below). However, very little is known about how mitotic-endocycle transitions are controlled at specific positions and cell types in a developmentally and physiologically dependent manner.

Root meristem maintenance

The root meristem maintenance is essential for the overall rate of root growth and meristem size. This is achieved by an appropriate balance between the production of new cells and subsequent expansion and differentiation. Endoreplication appears to be very important in the root meristem maintenance as an early onset of the endoreplication can have severe consequences on the growth and development. For instance, mutation of the SUMO E3 ligase *HPY2* results in stunted root growth because of an increase in cellular endoploidy at the expense of cell proliferation[212].

In *Arabidopsis*, two CCS52A isoforms CCS52A1 and CCS52A2 have been found that regulate the meristem size through different mechanisms[213]. CCS52A1 controls meristem size by stimulating endoreplication in the elongation zone and spatially determines the root-meristem-elongation zone border. The mutation in *CCS52A1* results in the longer roots containing more dividing cells in the meristem and delayed endoreplication. In contrast, CCS52A2 regulates the meristem maintenance and structure by repressing the mitotic activity in the QC and stem cells of the root meristem. The mutation in *CCS52A2* results in the emergence of differentiated root hair in close vicinity of the meristem, disturbs the regular differentiation patterns of root cell types, brings irregularities in the root size and shape and disorganises the stem cell niche by making QC and stem cells indistinguishable. The ICK3/KRP5 a CDK inhibitor that promotes cell elongation and endoreplication in the elongation zone, has been shown to have a rate limiting role for the primary root growth[214]. The loss of function of *KRP5* leads to smaller roots and reduced final cell sizes. In the embryonic root, it is also expressed in the transition zone between root and hypocotyl, the loss of function leads to delayed germination[214].

In addition, the meristem size is known to be developmentally regulated by the antagonising action of auxin and cytokinin as this action is associated with the mitotic-to-endocycle transition[205,215] (Figure 3.5). Auxin is known to inhibit the endocycle onset, whereas cytokinin promotes it[215]. In the elongation zone, cytokinin signalling appears to control endocycle entry via combined actions of B-type *Arabidopsis* Response Regulators (ARRs) ARR1, ARR12 and ARR2[216–218]. In case of ARR2, a recent study identified that it acts as a transcriptional activator of *CCS52A1* in the root meristem that in turn trigger the mitotic exit and onset of endocycle.

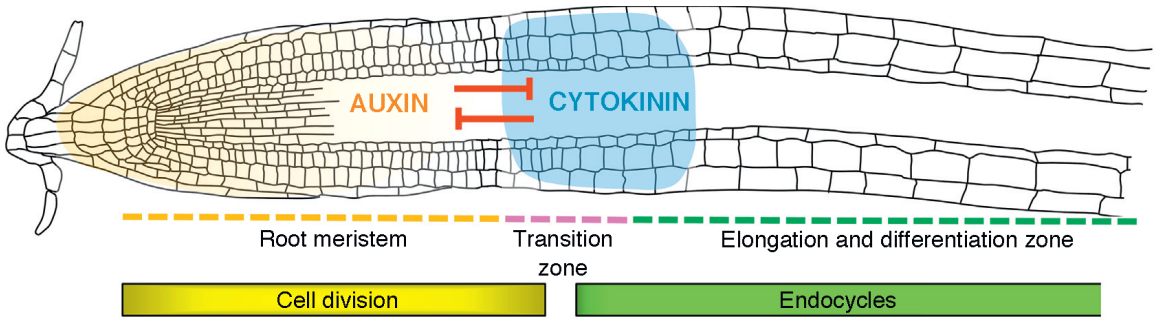


Figure 3.5: Endocycle during development of *Arabidopsis* root. The antagonising action of auxin and cytokinin developmentally determine meristem size and onset of cell elongation and differentiation in the root transition zone. Concomitantly with the onset of cell differentiation, cells transit into the endocycle programme. This figure is adapted from[205].

Collet hair elongation

Collet is the transition zone between the root and hypocotyl junction. The epidermal cells on the collet are thought to anchor the seedling to the substratum and facilitate geotropic responses and water uptake well before the actual root hairs have developed[219]. Though the collet epidermal hair development is not documented well, the collet initials can be seen as early as the heart stage of embryo development and become clearly demarcated during seed germination. Until post-germination, collet epidermal cells change in length by 2 fold and elongate before they forms hairs[220]. Recently, it has been shown that the hairs develop simultaneously from all epidermal cells of the collet and the endoploidy levels increases during their development from 4C before bulge formation to 16C by the time that hairs achieve their full length[221]. However, so far no mutant study is available that couple collet hair elongation with the endocycle.

Root hair development

Root hairs play an important role in water and nutrient uptake. Epidermal cells produced in root apical meristem become hair cell or non-hair cell in the differentiation zone based on their position relative to the underlying cortex cells[211]. An epidermal cell in contact with two cortex cells becomes hair cell while that in contact with one cell develops into a non-hair cells. Previous studies identified that a group of components are involved in the specification and differentiation of hair and non-hair cells[211,222]. Specifically in non-hair cells TTG1/GL3/EGL3/WER activates the transcription of *GL2* and *CPC*. Activation of GL2 results in non-hair cell fate while CPC travels into the presumptive hair cells, where it competes with WER for binding to the TTG1/GL3/EGL3 complex during which WER gets repressed by the SCM allowing CPC to over-competes with WER, resulting in loss of *GL2* activation and consequently hair cell specification[211]. Still, it is not clear whether (like trichomes) the root hair fate is linked with entry into an endoreplication cycle or not. However, root hairs eventually enter an endoreplication cycle and increase the size of their genome by up to 16C. Recent study indicates that root hair elongate independently of the endocycle, which is controlled by newly discovered *BHLH* transcription factor, *RSL4*[223]. The *rsl4* mutants display severe defects in root hair growth, whereas constitutive expression of 35S:*RSL4* leads to continuous growth of hair cells without increase in the nuclear content.

3.3 Environmental and hormonal control of the endocycle

In plants, different environmental factors exert control on endoreplication levels of cell types, tissues and organs[208]. Figure 3.6 gives an overview of the endoreplication extent under such environmental factors. Light is one of the important environmental factors known to impact the endoreplication levels of plant organs. In the *Arabidopsis* hypocotyl, the third endocycle is inhibited by light through the action of the red/far red light photoreceptor phytochrome[184] (see Section 3.2.1). In wild type plants during photomorphogenesis (growth in light), up to two rounds of endoreduplication are observed in hypocotyl cells, whereas a third round takes place only during skotomorphogenesis (growth in dark). In contrast to wild type plants, phytochrome A mutants also have 16C nuclei in hypocotyl cells when grown under continuous far red light, suggesting the importance of phytochrome in the repression of endoreplication. In contrast to increased endoploidy in hypocotyl during growth in dark, a partial shading decreases endoploidy in *Arabidopsis* leaves[224] suggesting that light might have opposite effects depending on the organs.

Water availability appears to be another environmental factor that influences endocycle. In the wild-type *Arabidopsis* leaves, water deficiency treatment was shown to reduce cell size and the extent of endoreplication[224]. In contrast to wild-type, the transgenic plants with an increased endoreplication level were less sensitive to the stress due to a higher leaf expansion rate and maintenance of cell size under water deficiency[224]. This suggest that, an increased endoploidy level confer advantage under water deficit conditions. Similarly, endoreplication has seen to yield an advantage under UV-B stress conditions. In the wild-type *Arabidopsis* plants, acute UV-B treatment is known to provoke a decrease in both cell number and the average cell size. The mutant plants for the endocycle regulator *E2Fe/DEL1* showed resistance to the treatment and displayed smaller reduction in leaf size area than that in control plants. Correspondingly, the mutant showed an elevated percentage of high-ploidy cells (8C and 16C), which suggests that the mutant plants probably use the growth potential stored in their polyploid cells to compensate for the decreased cell number[208].

Plants respond to genotoxic stress, which causes DNA double strand breaks, by inducing endoreplication in root tips[225]. As polyploid cells rarely divide, the induction of endoreplication after DNA damage has been speculated as a mechanism to prevent the transmission of DNA lesions into the pool of meristematic cells by pushing the damaged cell into a non-dividing state, thus safeguarding the progeny from DNA mutations[225]. Similarly, the DNA damage caused by oxidative stress stimulates endoreplication in *Arabidopsis* plants. Osmotic stress in *Arabidopsis* leaf has shown to induce endoreplication earlier than the control plants[226]. This response was shown to be mediated by gibberellic acid (GA)–DELLA pathway. Like in osmotic stress, salt stress results in a decrease of active GAs, which in turn stabilises DELLA proteins to repress cell division and post-mitotic cell growth[227]. This suggests that salt stress might induce endoreplication as well.

Extreme cold and heat treatments in plants have shown to reduce the extent of endoreplication. For instance, the root cortex and hair cells of chill treated (10°C) soya bean seedlings showed very low proportions of 8C and 16C nuclei compared to the control (25°C)[228]. Moreover, different species and organs respond differently towards the temperature ranges and durations of the treatments. Short term high temperature treatment does not affect endoreduplication in maize endosperm, but prolonged

high temperature treatment results in reduced endoploidy levels[229]. Nevertheless, a mild increase in growth temperature had a positive effect on endoreduplication cycles in tomato pericarp cells[230]. Low temperature treatment (15°C versus 25°C) decreases endoreplication onset and growth rates in orchid *Oncidium varicosum* flowers[231]. In *Arabidopsis*, a negative correlation was found between temperature and endoploidy levels. Plants grown at low temperature (18°C and 14°C versus 22°C; 16-h light:8-h dark in growth chamber) showed significant (but slight) increase in the endopolyploidy levels[232].

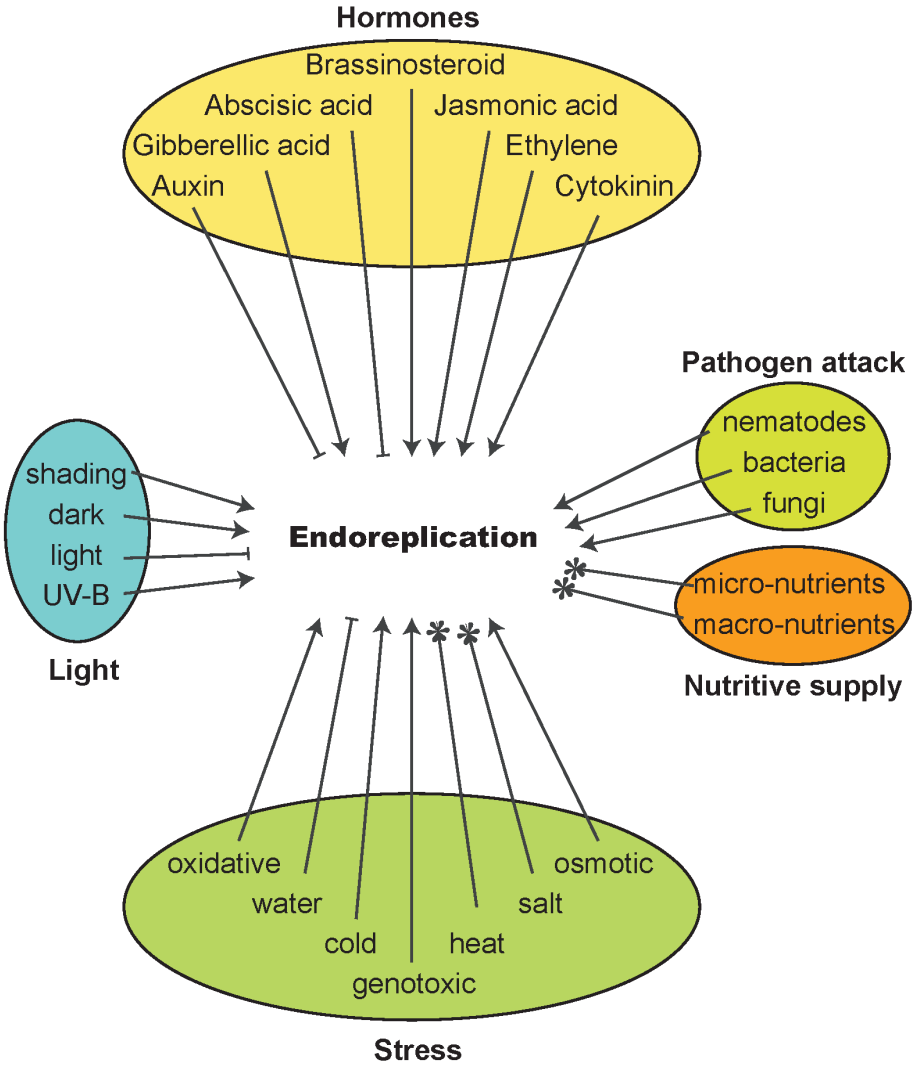


Figure 3.6: The extent of endoreplication in response to different environmental conditions. The pointed arrow head indicates the stimulation of endoreplication, whereas blunt indicates the suppression. The snowflake arrow head indicates that the extent of endoreplication under particular conditions is non-informative. This figure indicates a compilation of reported effects in different organs of Arabidopsis thaliana plants.

Pathogens attacks e.g. by powdery mildew[233], nematodes[234,235] and bacteria[236] are also known to trigger endoreplication at the interaction sites. In addition, symbiotic interaction with arbuscular mycorrhizal fungi[237], nitrogen-fixing bacteria[159,238] increases endoploidy levels in the infected host cells. Recent analyses have shown that induction of endoreplication by symbiotic fungi is a widespread phenomenon and can be found in many angiosperm groups[239]. The impact of both macronutrients such as phosphate, sulphur, nitrates, etc. and micronutrients such as iron, boron, etc. deprivation on endoreplication is so far insufficiently investigated. In one such study, the plants grown under standard potting soil versus sandy

substrate showed no significant impact on endoreplication levels[232]. However, those results are to be taken with caution, as other studies reported that endopolyploid species have higher nutritive demands than most species without endopolyploidy, and are found preferentially in habitats that require a fast completion of growth and fast development supported by optimal nutrient supply, which suggest that a high DNA content plant may require a richer food supply[240].

Environmentally modulated endoreplication is likely to be mediated by phytohormones, e.g. ethylene biosynthesis upon UV-B treatment has been hypothesised to mediate an increase in DNA endoploidy levels of trichome socket cells of cucumber *Cucumis sativus* cotyledons. Concomitantly, treatment of *Arabidopsis* hypocotyls with the ethylene precursor 1-aminocyclopropane-1-carboxylic acid induced an additional endocycle, whereas ethylene-insensitive mutants showed a slight decrease in endoploidy levels[241]. Gibberellic acid treatment in wild type *Arabidopsis* hypocotyls has shown to induce endoreplication, whereas GA deficient mutants show decreased endoploidy levels in hypocotyls[242]. Similarly, mutation in *SPY*, a negative regulator of gibberellin signalling, showed increase in endoploidy of *Arabidopsis* trichome[243]. In contrast to positive effects of ethylene and gibberellin, auxin has been shown to negatively regulate endoreplication in *Arabidopsis* root tips[215]. Antagonistic to auxin, cytokinin has been shown to positively regulate endocycle in *Arabidopsis* root[215].

Most of the aforementioned studies have been mainly carried out on leaves and hypocotyls as preferred organs and identified that different environmental factors exert different controls on endoreplication levels. In addition, species and organs may have different responses to such environmental factors. Nevertheless, very little is known about how such environmental factors control endoreplication levels in intact (as a whole organ) root as well as individual tissue types.

3.4 Modeling approaches for endocycle

During mitotic cell cycle to endocycle transition, the M phase is skipped without blocking S phase. In plants, similar to animal cells, this transition is achieved by down-regulating the mitotic CDK that is required for G2-M transition, while simultaneously allowing continued activity of the S phase CDK that drives G1-S transition. The underlying mechanisms of these processes are very complex and vary widely between cell types and organisms. In recent years, several predictive models have been used to understand the operating principles of complex regulatory networks involved in mitotic cell cycle exit[244–246], endocycle onset and endocycle progression[247,248]. Such models have been beneficial in identifying the main regulators involved in these processes and designing further experiments. For example, using a mathematical model of the *Arabidopsis* endocycle, cyclic accumulation of KRPs that periodically inhibit S phase CDK activity, combined with specific inhibition at Mitotic CDK activity by SIM, has recently been proposed to be sufficient to promote endoreplication[248]. Similarly, a mathematical model demonstrated that cyclic accumulation of the transcription factor E2f1 (E2f–FlyBase) is essential for endoreplication in the highly polyploid Drosophila salivary gland[247].

Beside models focusing on gene regulatory networks, recently an approach was used to answer a long standing question in plant development, which is to know what comes first, endoreplication or cell elongation[249,250]. This approach used a combination of DNA replication imaging and optical estimation of

the amount of DNA in each nucleus to determine the boundary region between the meristematic and elongation zones in a developing *Arabidopsis* root and identified that endoreplication precedes rapid cell elongation in root[251]. Despite these efforts, the cellular arrangement of such dividing and endocycling cells in a developing organ is still missing. Moreover, it would be interesting to see the extent and order of endoreplication among different tissue types in an organ under developmental as well as environmental cues.

3.5 Author contributions

I wrote this chapter by myself. It resulted from the many fruitful discussions with both my promoters.

Chapter 4

A spatiotemporal DNA endoploidy map of the *Arabidopsis* root reveals a role of the endocycle in stress adaptation

Rahul Bhosale*, Veronique Boudolf*, Gert Van Isterdael, Georgina M. Lambert, Ilse Vercauteren, Fabiola Cuevas, David W. Galbraith, Steven Maere#, and Lieven De Veylder# (In Preparation For: *Nature*).
*These authors contributed equally to this work, #Shared corresponding authors

Abstract

Endoreplication, a variant cell cycle process that results in endoploidy due to genome duplication in the absence of mitosis or cytokinesis, is observed in arthropods, molluscs, and vertebrates, but is especially prominent in higher plants where it is essential for cell growth and fate maintenance. However, a comprehensive view on the physiological significance of the endocycle remains elusive. Here, we reverse engineered and experimentally verified a high-resolution DNA endoploidy map of the developing *Arabidopsis thaliana* root, revealing a remarkable spatial and temporal control of DNA endoploidy distribution across tissues. Our virtual root endoreplication model allows accurate prediction of DNA endoploidy changes in response to perturbations, and reveals a strong dependence of the endoploidy distribution on stress signals. For instance, root measurements of endoreplication mutants grown under salt stress demonstrate a role for the endocycle in rapid adaptation to salinity. Combined with the observation that endopolyploidy occurs most frequently in plant species grown under extreme or variable conditions, our data might help explain the widespread occurrence of the endocycle across dicots.

For the author contributions, see page 91.

4.1 Background

Endoreplication is a specialised mode of cell cycle during which cells undergo extra rounds of DNA replication without intervening cell divisions, and it is often closely associated with specific cell types, organs, and developmental stages[142,185]. In animals, endoreplication has a recognised role in driving body size[168] or in maintaining tissue and organ growth primarily as a part of their developmental programme[183] and to lesser extent in response to exogenous stresses such as regeneration of damaged liver and cardiomyocytes[191]. In plants, endoploidy is frequently observed as an essential aspect of cell growth[185,186,249] and differentiation[206] as well as a prominent response to stress conditions[208]. Light[147,184], DNA damage[225], pathogen attack[233–236] and other stress conditions such as drought[252], temperature[232], etc. have been shown to activate endoreplication in plants. Due to a plant's immobile life style, stress-induced endoreplication has been postulated to be a mechanism facilitating adaptive growth[225,240,253].

Although over recent years many genes have been identified that control endoreplication onset and progression, the physiological role of the endocycle in sustaining plant growth in response to stress has been poorly characterised due to lack of a clear knowledge on the temporal and spatial occurrence of endoreplication. A major open question is how cells with different DNA endoploidy levels are integrated into a developing organ, and how this organization contributes to the growth of the plant under different environmental conditions. To understand endoreplication in a spatiotemporal context, we constructed a DNA endoploidy map of the developing *Arabidopsis* root, which displays a simple radial symmetry, with one-cell-layer cylinders of epidermis-, cortex-, endodermis-, and pericycle cells surrounding the vascular bundle. Within the root, all cells arise linearly from a group of stem cells surrounding the quiescent center (QC). Close to the QC, cells are dividing. As cells age, they gradually lose their division competence, and enter the endocycle, resulting in cells having a 4C, 8C, or 16C DNA content.

4.2 Results and discussion

4.2.1 Endoploidy-enriched transcripts show association with ST root organisation

To assess endoreplication-dependent gene expression levels, *Arabidopsis thaliana* root cortical nuclei from *pCO2:YFP-H2b* line[254] (Col-0 ecotype) were flow sorted on their DNA content (2C, 4C, 8C and 16C) and employed for endoploidy-specific transcriptome analysis (Method Section 4.6.3). In total, 3,737 genes were differentially expressed (P value < 0.05, Benjamini-Hochberg FDR correction) across at least two endoploidy levels (Supplemental Data Set 1). These were grouped into 24 clusters (i.e. the total number of possible expression level rank patterns over the four endoploidy levels; Supplemental Data Set 2), exhibiting various patterns of endoreplication-dependent expression. The centroid pattern of each of these clusters is represented in Figure 4.1. These clusters were further classified as endoploidy-specific based on their peak expression endoploidy level. Analysis of the functional enrichment (Supplemental Data Set 3, Figure 4.2) and spatiotemporal (ST) peak expression (Supplemental Data Sets 4 and 5, Figure 4.3) of these clusters revealed that 2C and 4C specific transcripts, enriched for DNA replication

and cell wall biogenesis related genes, are predominantly expressed in the developing xylem, phloem, atrichoblast and endodermal cells within the meristematic zone of the root expression map reported by Brady and co-workers[2]. In contrast, 8C and 16C specific transcripts, enriched for genes involved in stress responses and trichoblast differentiation, are primarily expressed in the transition and maturation zones of cortex, endodermis and hair cells.

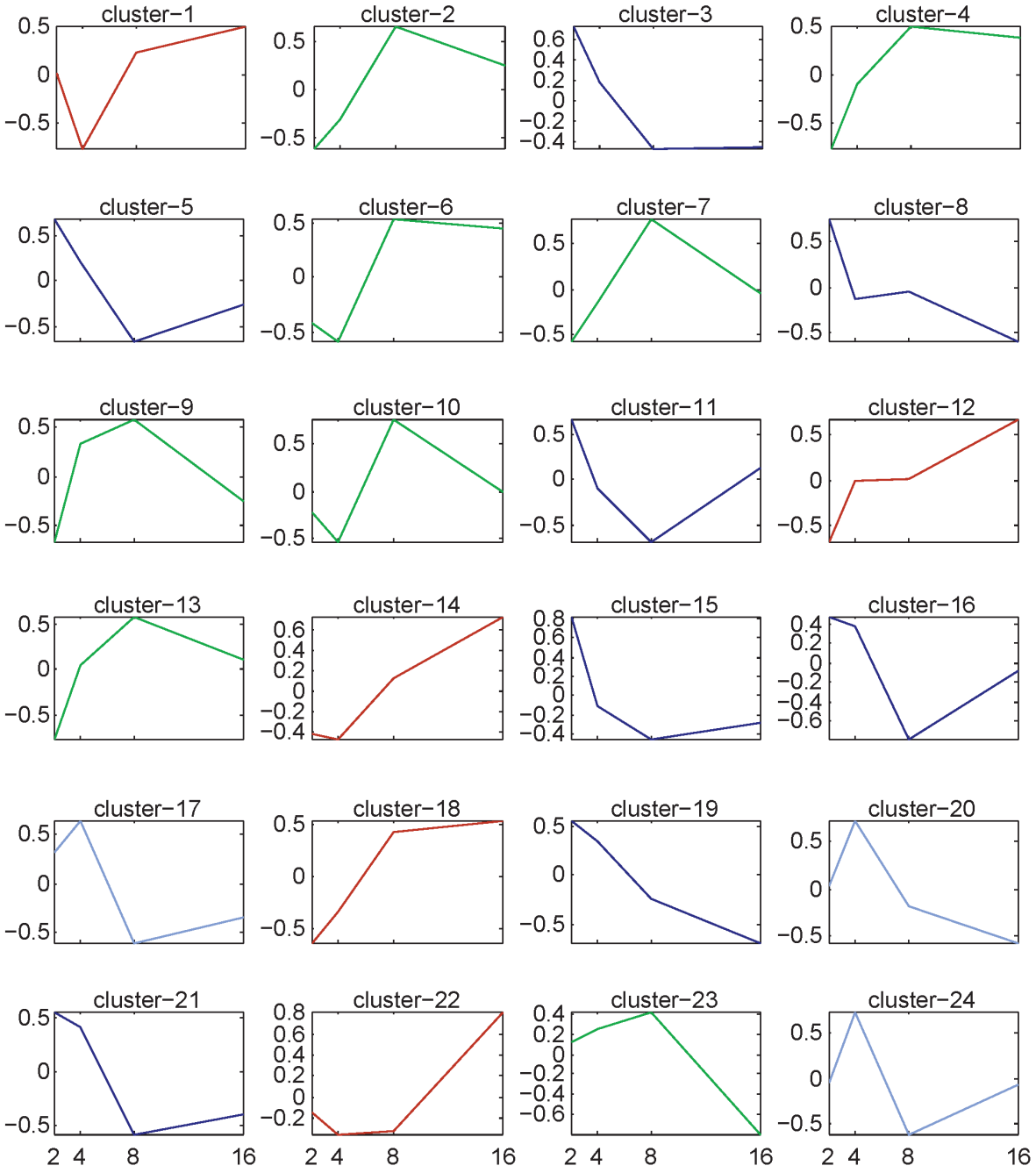


Figure 4.1: Centroid patterns and endoploidy-specific classification of 24 clusters. Differentially expressed 3,737 genes were k-mean clustered using matlab function 'k-mean' into 24 possible expression level patterns. These clusters are further classified into four endoploidy classes i.e. 2C (dark blue), 4C (light blue), 8C (green) and 16C (red line plots) based on the peak expression of the endoploidy.

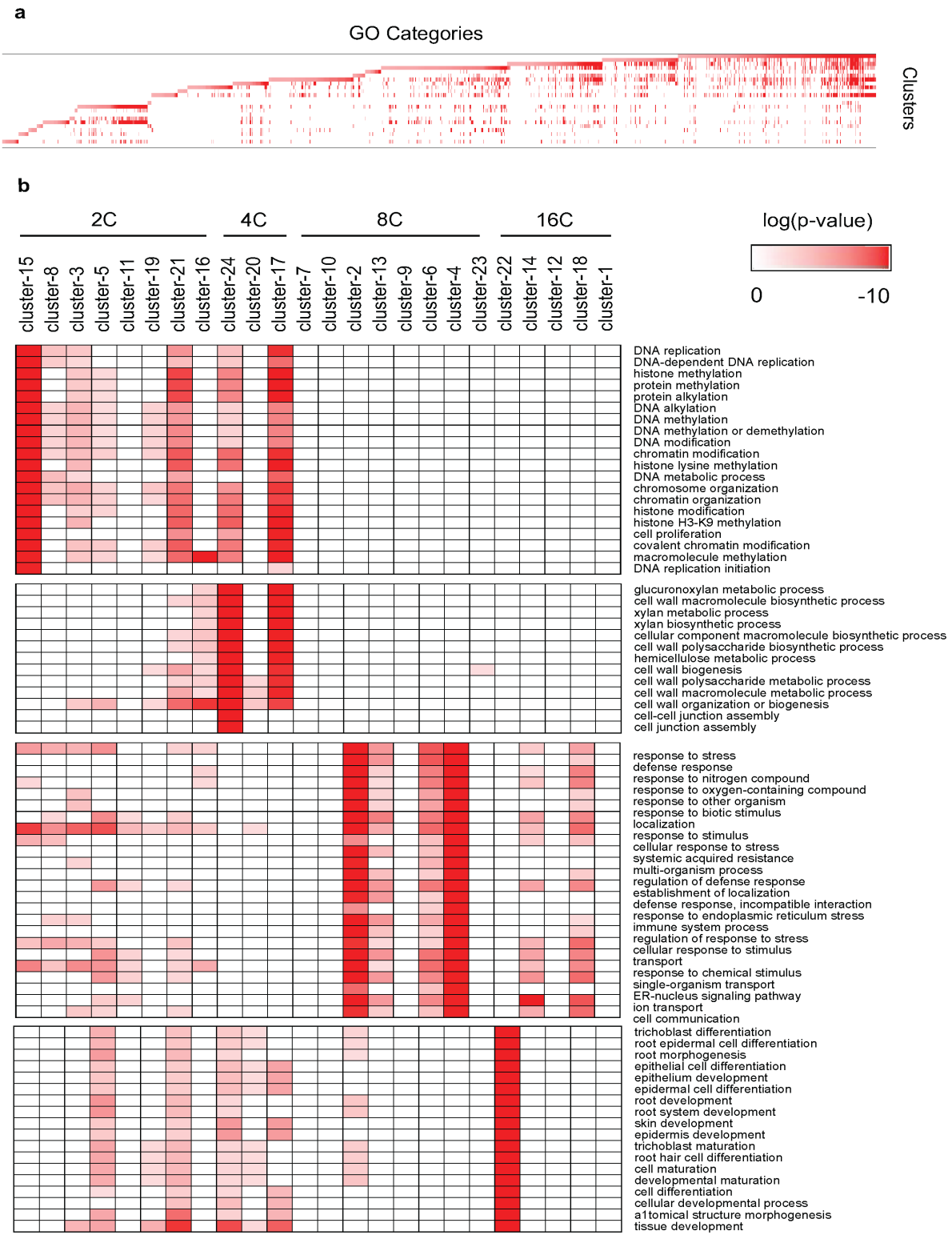


Figure 4.2: Functional enrichment of 24 endoploidy-specific expression data clusters. (**a.**) Hierarchical clustering of the enriched GO term profiles across the 24 endoploidy-specific expression data clusters. Functional enrichment was calculated with the BiNGO tool[40] using hypergeometric tests and Benjamini-Hochberg correction at FDR = 0.05 (**b.**) Pruned version of **a** showing only the top enriched GO terms (P value <1*E-10) derived from one example cluster for each individual endoploidy (2C, cluster-15; 4C, cluster-24; 8C, cluster-4 and 16C, cluster-22). The 2C and 4C specific transcripts are enriched for DNA replication and cell wall biogenesis processes, while 8C and 16C specific transcripts are enriched for genes involved in stress responses and root hair differentiation, respectively. The enrichment analysis suggests that endocycle is linked with the programmes underlying root development.

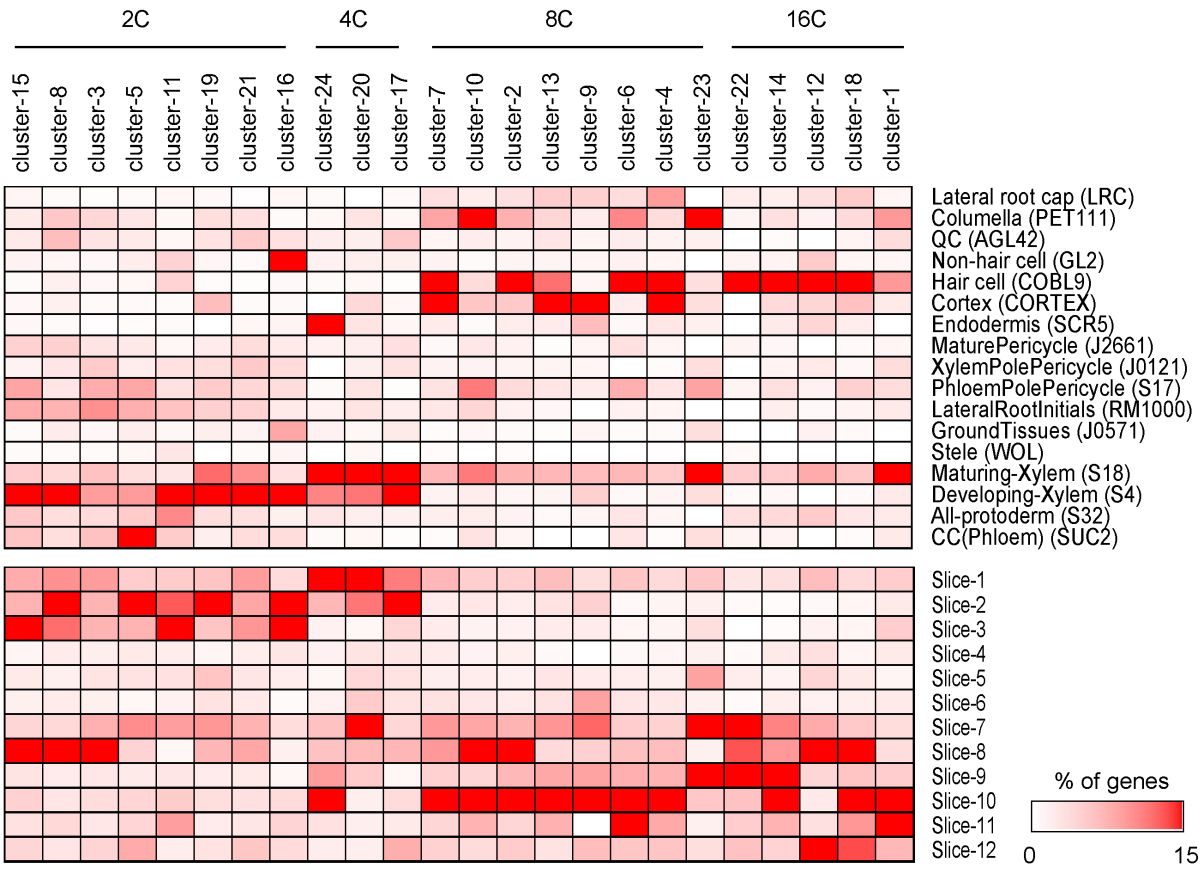


Figure 4.3: Peak expression distribution of cluster genes in the root expression map[2]. Displayed is the proportion (%) of transcripts that are peak-expressed in any of 17 marker lines covering one of 14 root tissues or a combination of tissues (top) and any of 12 developmental stages (slices, bottom), of the *Arabidopsis* root. The tissue coverage of the marker lines (indicated in parentheses) is as described in Brady et al. (2007)[2]. The 2C and 4C specific transcripts are mainly enriched in inner tissues (xylem, phloem, and endodermis) and in early developmental stages (Slice 1-3), while 8C and 16C specific transcripts are primarily enriched in outer tissues(cortex, epidermis and lateral root cap) and in later developmental stages (slice 7-12), suggesting spatio-temporal association of endoploidy levels with root organisation.

4.2.2 A predicted endoploidy map reveals spatial and temporal control of DNA endoploidy distributions across tissues

The observed association of endoreplication-enriched transcripts with specific ST root zones (Figure 4.3) suggest that endoploidy-dependent gene expression levels could be used to predict the nuclear DNA content status of distinct root tissues at different developmental stages. To this end, we constructed a mathematical model that predicts the expression level of genes in 12 different root slices and 14 different tissues (measured by Brady and co-workers[2]) as a function of their endoploidy-specific expression levels in the cortex (see Methods Section 4.4.1, Mathematical Model I, Figure 4.4). The model assumes that the endoploidy levels in cells along longitudinal axis follow logical increments (2C->4C->8C->16C). Briefly, the expression level of a gene in a particular slice is assumed to be the sum of the gene's expression levels in all cells of different types (and possibly with different endoploidy levels) within that slice and the expression level of a gene in a particular marker line or tissue is the sum of the gene's expression levels in all marked or single tissue cells, respectively, which may have different endoploidy levels (Figure 4.4a). The parameters in the model are the nuclear DNA content boundaries along the longitudinal axis of the different root tissues(Figure 4.4b), the position of which is optimised to obtain the best possible fit between

the predicted gene expression levels and the levels measured by high-resolution root expression profiling experiments[2] (Figure 4.4d; Figure 4.5). Overall, the model determines the fraction (normalised values between 0 and 1) of 2C, 4C, 8C and 16C cells (using adopted ST cell count matrix reported by Cartwright and co-workers[255]; See Section 4.4.1 and Table 4.1) in a slice or tissue to calculate the contribution of each endoploidy-specific expression towards the total expression of a particular gene in that slice or tissue.

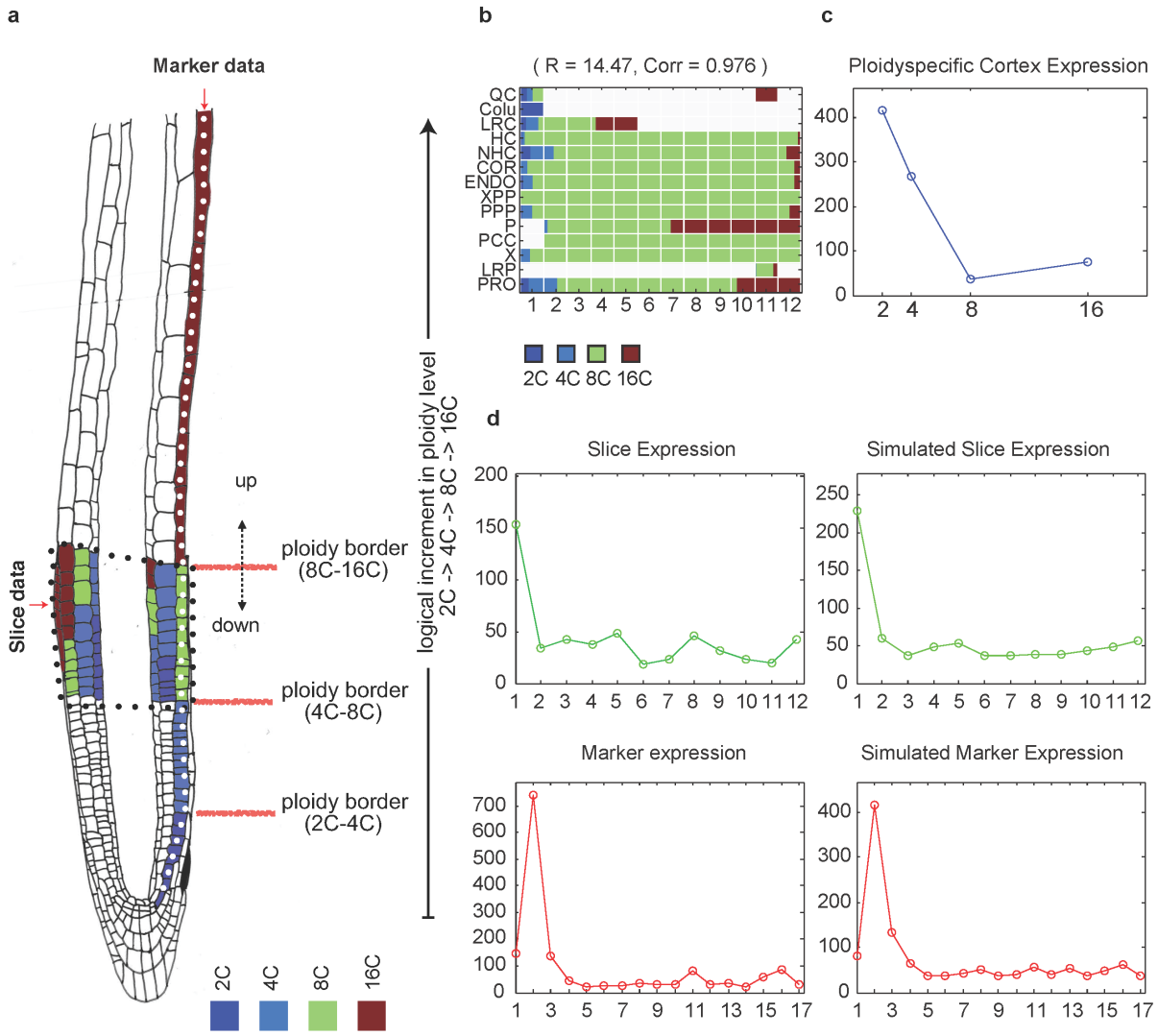


Figure 4.4: Schematic representation of assumptions used in the model and example of the optimised expression patterns and endoploidy map learned for one gene using Mathematical Model I. (a.) The model assumes that the endoploidy levels in cells along longitudinal axis follow logical increments (2C->4C->8C->16C). The endoploidy boundaries are the parameters in the model and moved one cell in a single cell type either up or down during the optimisation process. The marker line expression level of a particular gene is a sum of the expression levels of a mixture of cells from a single cell type (white dots) with different endoploidy levels (2C-16C, see legend). The slice expression level is a sum of the expression levels of a mixture of cells (area under black dots) from multiple cell types with different endoploidy levels. **(b.)** The endoploidy map of the *Arabidopsis* root learned from a single gene (AT3G50240). **(c.)** The endoploidy-specific expression pattern of the gene in **b**. **(d.)** The green and red line plots indicate the measured (left, from Brady and co-workers[2]) and simulated (right) expression patterns of the gene in a in slices and marker lines, respectively.

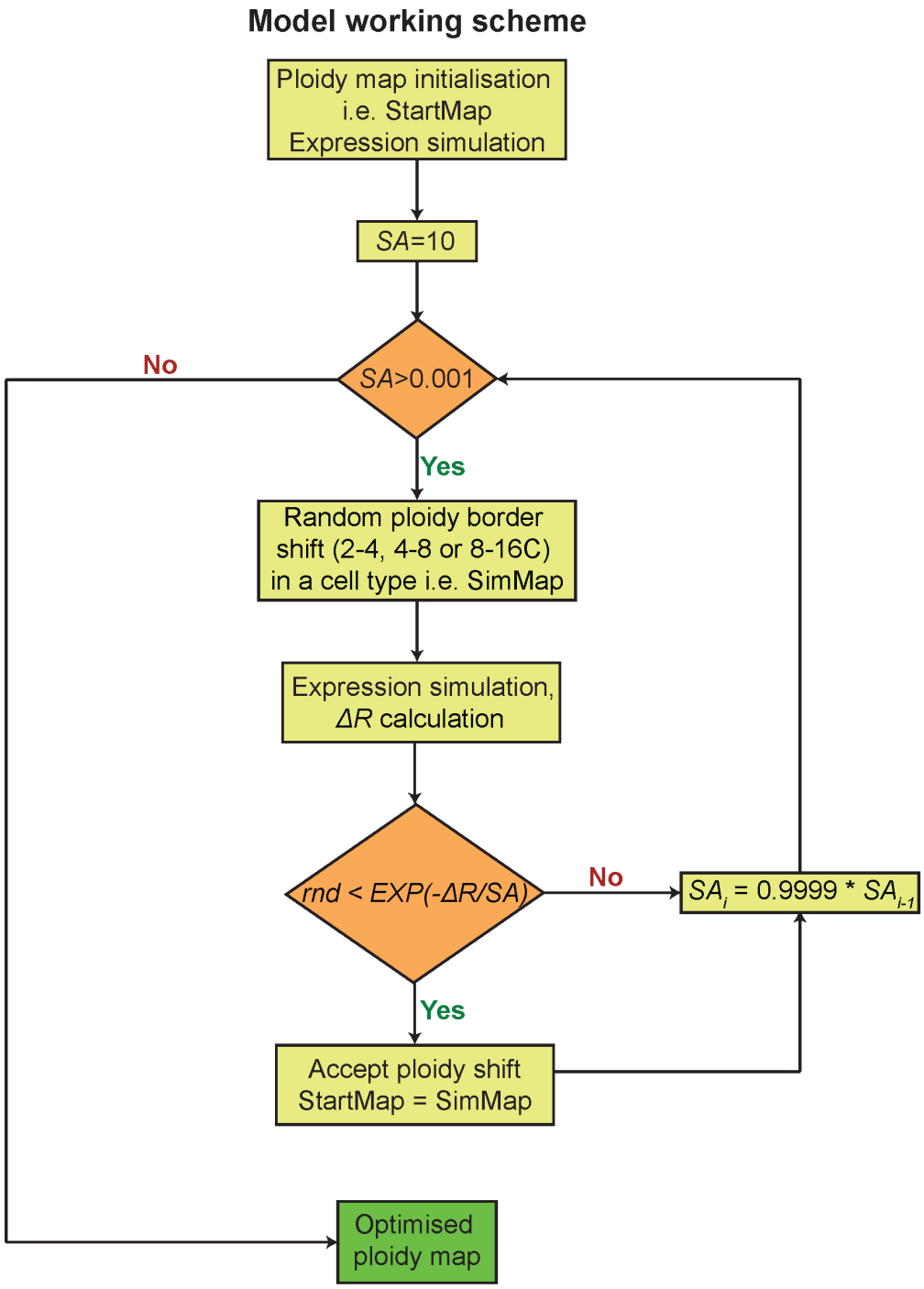


Figure 4.5: Schematic representation of the mathematical modeling approach. Initially, the model randomly positions the endoploidy boundaries on the map. At each optimisation step, a random endoploidy boundary shift is performed, after which the expression levels of the modelled genes are simulated in all slices and cell types based on the changed endoploidy map. A step is accepted or rejected based on the Simulated Annealing criterion in Equation 4. Then the SA temperature is lowered and a new step in parameter space is attempted. Over the course of an optimisation run, the location of the endoploidy boundaries is gradually adjusted to optimise the fit of the modelled gene expression levels to the levels measured by Brady and co-workers[2]. Eventually, the optimised endoploidy map with the lowest chi-squared (R) score is obtained.

Since gene expression may be regulated in a tissue- and developmental stage-specific manner, it is not expected that every gene's spatiotemporal expression pattern can be predicted accurately from its endoploidy-specific expression pattern in the cortex. In other words, not all spatiotemporal gene expression profiles are adequately (i.e. exclusively or to a large enough extent) reflecting endoploidy-specific gene expression changes for the purpose of reconstructing a endoploidy map of the developing

root. Thus, we used a feature selection (See Method Section 4.5.1) approach to obtain a set of 332 marker genes that were able to reliably predict their ST expression patterns from their endoploidy-specific expression pattern in the cortex. Combining the endoploidy predictions of a balanced set of 332 endoploidy markers (Supplemental Data Set 6), a reproducible and stable DNA endoploidy map (in terms of model convergence across different runs) was obtained for the complete root (Figure 4.6).

This map reveals clear differences in the endoploidy distribution over the distinct tissues, suggesting that endoreplication is under strict spatiotemporal control. The endoploidy levels across tissues appear to correlate primarily with the radial organisation of the root, with the outermost tissue layers displaying a higher endoreplication level than the inner tissue layers. Remarkably, a clear difference was observed between the different epidermal cell layers, with hair cells undergoing a third endocycle earlier than non-hair cells. Among vasculature tissues, xylem appears to undergo endoreplication earlier than other tissues. Among non-vascular tissues, phloem pole-associated pericycle cells appear to undergo endoreplication in late developmental stages and xylem pole-associated pericycle cells appear to be mainly dividing. This observation corresponds with the concept of an 'extended meristem' i.e. the xylem pole-associated pericycle cells acts as stem cells to retain the capacity to undergo (asymmetric) cell division higher up in the root when other cells have differentiated and undergo lateral root initiation. This allows the root to have high-flexibility to respond to an ever-changing environment in the soil.[256].

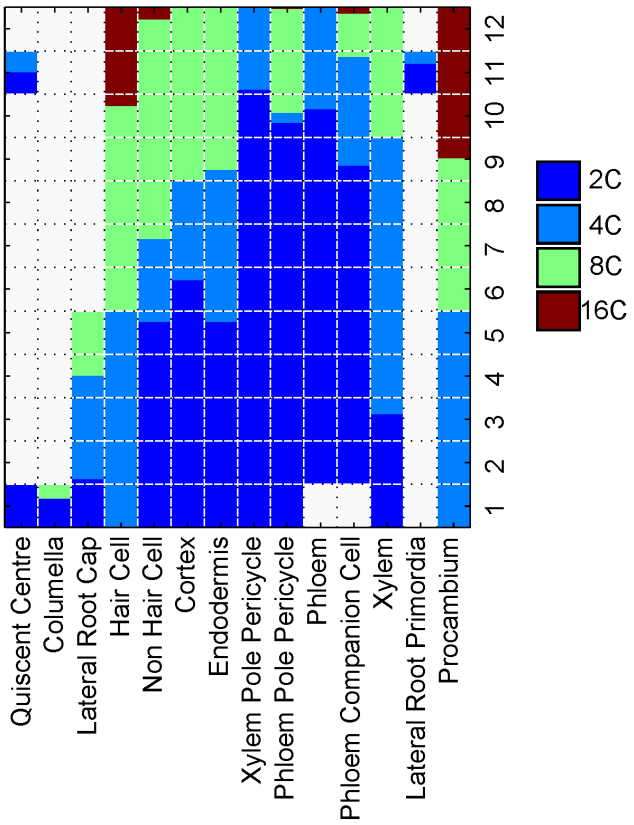


Figure 4.6: Predicted root endoploidy map. The endoploidy map of *Arabidopsis* root learned using set of 332 representative genes in which columns indicate the tissue types and rows indicate the developmental stages (Slices 1-12 in root expression map[2]). The outermost tissue layers (epidermis, cortex and endodermis) are predicted to have higher endoreplication levels than the innermost tissue layers (Pericycle, Xylem and Phloem), suggesting a correlation with the radial organisation of the root.

4.2.3 Experimental validation confirms the reliability of the predicted endoploidy borders

The simulated endoploidy map was experimentally validated by measuring, via flow cytometry, the DNA content of cells within specific tissues in root tips, using cell type-specific GFP marker lines (Methods Section 4.5.1, Figure 4.7). The measured endoploidy distributions were scaled to fit the tissue- and stage-specific cell counts underlying the virtual endoploidy map (Figure 4.8a). The experimental and predicted maps present a very similar overall picture of spatiotemporal root nuclear DNA content organization (Figure 4.8b), except for the location of the 2C-4C boundaries, likely due to the fact that flow cytometry cannot distinguish G2 non-endoreplicating nuclei from G1 endoreplicating nuclei, both of which have a 4C DNA content. Also, the endoploidy boundary positions observed in the validated map are not always reliable, as they are highly dependent on the absolute number of nuclei or protoplasts extracted and measured per root in the flow cytometer analysis. The efficiency of protoplasting and nuclear extraction is dependent on the duration with which tissues are treated with the respective extraction buffers. Such dependency often gives quantitative variations in the endoploidy distributions, and it is not straightforward to compare data obtained from different experiments.

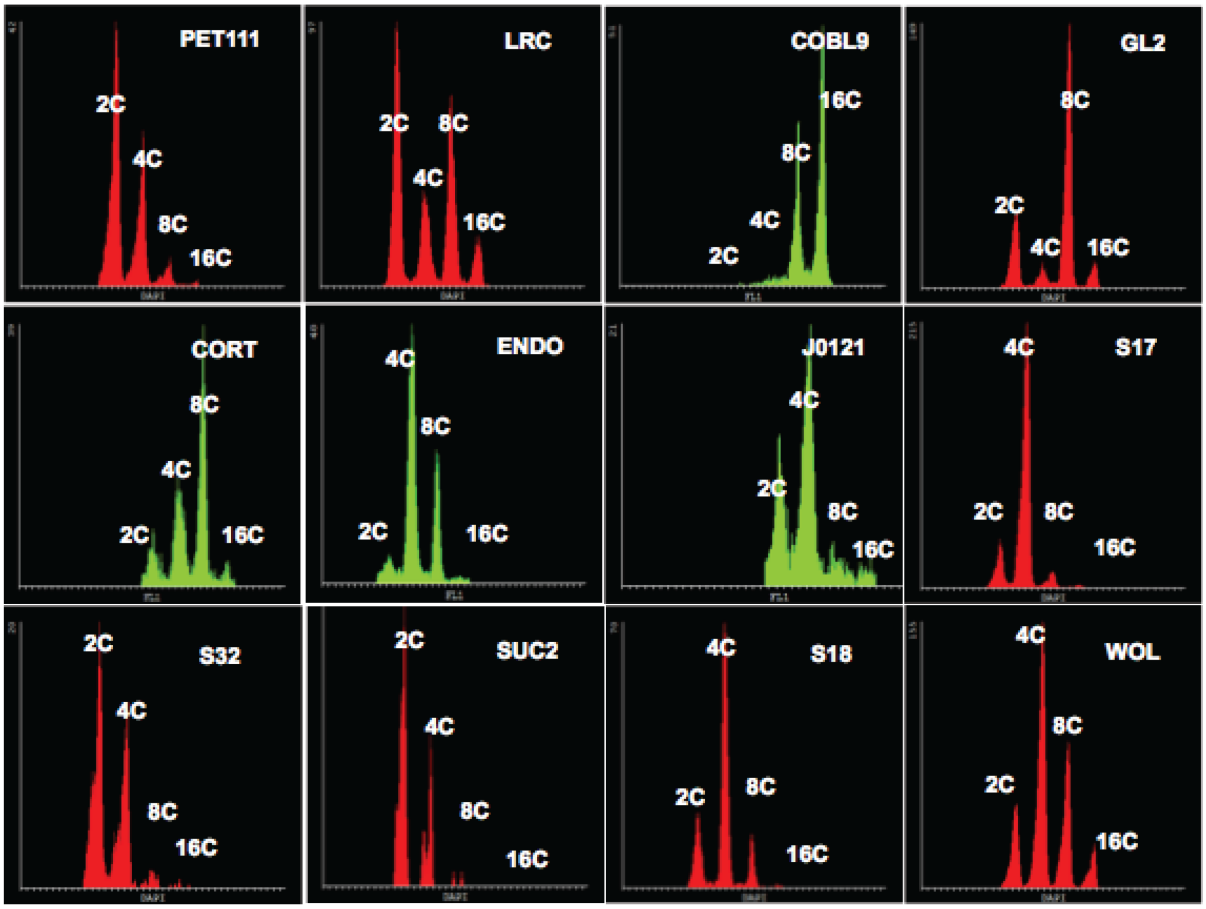


Figure 4.7: Endoploidy content profiles of marker lines (Table 4.2) obtained using flow cytometer analysis. Red histograms depict the profiles of marker lines subjected to protoplasting, while green histograms are obtained after nuclear extraction. The cut root tips of five-day-old plants of cytoplasmic and nuclear marker lines were treated with protoplasting solution and nuclear extraction buffer, respectively. The FACS sorted GFP-expressing protoplasts and extracted nuclei were stained with DAPI and measured using a CyFlow Flow Cytometer (Partec).

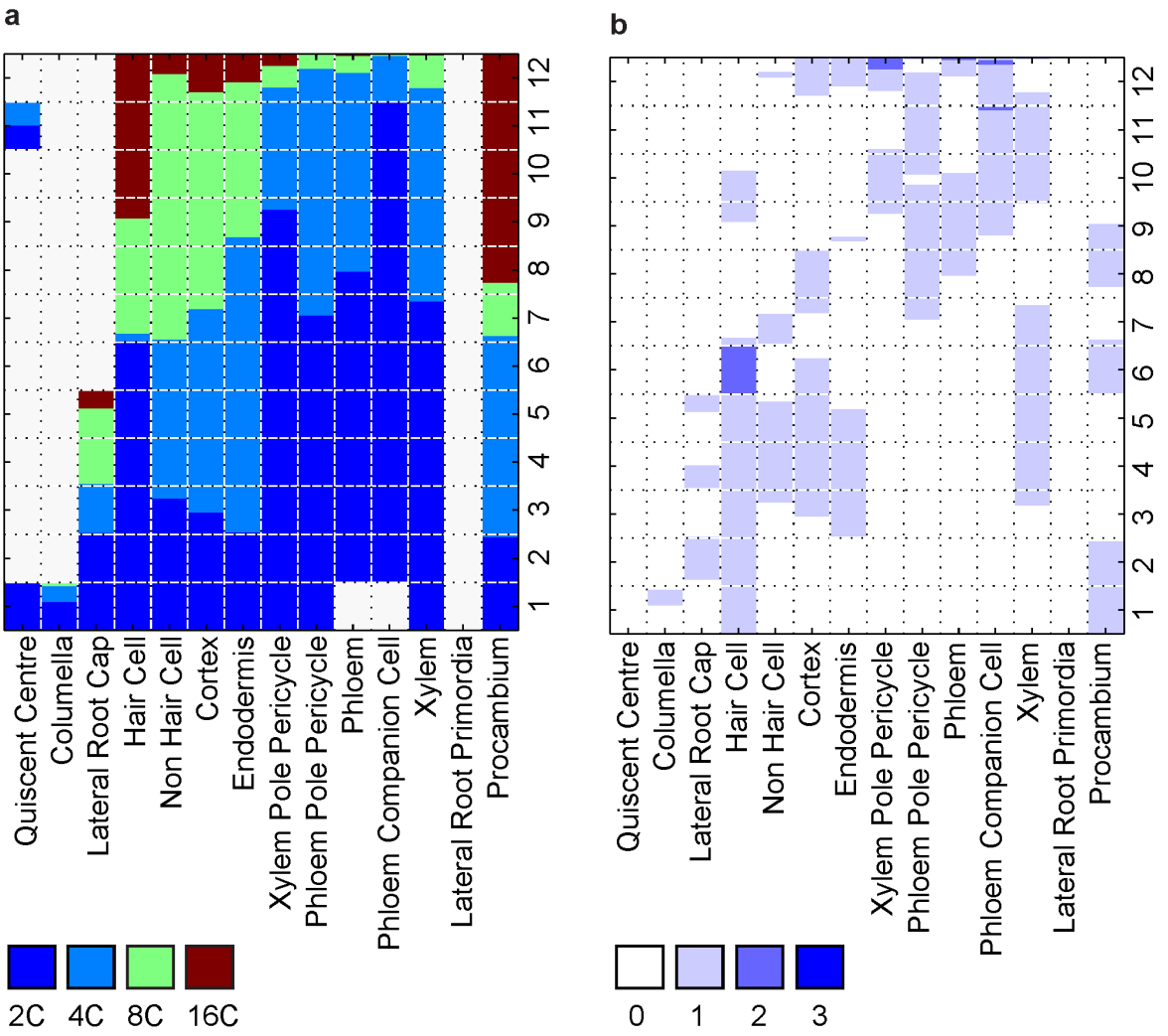


Figure 4.8: The validated map and its comparison with the predicted map. (**a.**) The map determined through flow cytometry analysis. The columns represent the 14 tissues and rows represent the 12 slices. The measured endoploidy distributions (in Figure 4.7) were scaled to fit the tissue- and stage-specific cell counts underlying the virtual endoploidy map.(**b.**)The difference between the endoploidy levels for each cell on the predicted map and the map determined through flow cytometry. The rows represent the tissue types and the columns represent the 12 developmental stages. The color indicates the difference between rounds of endocycle predicted vs. validated through flow cytometry at each cellular position on the map. The experimental and predicted maps show differences mainly in the location of 2C-4C boundaries.

Therefore, to reliably (independently of measured endoploidy profiles through flow cytometry) assess the location of 2C-4C endoreplication boundaries, we mapped the expression profile of the *SMR1* and *CCS52A1* endocycle-onset marker genes[213,257] (Figure 4.9a). First signals from the *SMR1:GFP-GUS* reporter were observed in the 19th ($\pm$ 1, n=6) atrichoblast cell, coinciding precisely with the 2C-4C border position on the predicted map. A weaker signal detected at the 29th ($\pm$ 1, n=6) cortex cell might mark the corresponding 4C-8C boundary. First signals from the *CCS52A1:GFP-GUS* reporter were seen in the 17th ($\pm$ 1, n=4) trichoblast cell, close to the end of the predicted 4C region, and in the 23rd ($\pm$ 1, n=4) cortex cell, which coincides precisely with the 2C-4C border position predicted by the map. Here, the first signal in trichoblast cell corresponds to the end of 4C region instead of the start, which might be explained by the fact that the trichoblast marker line COBL9 used in the model does not cover the first six slices thus the predictions for those slices can not be considered reliable. To validate the predicted order of endoreplication onset across different cell types, we examined sequential cross-sections of *SMR1* and *SIM* endocycle onset marker lines, and confirmed that endocycle onset in xylem cells precedes that of

atrichoblast and cortex cells (Figure 4.9b, c), whereas trichoblast cells engage into the endocycle before cortex and phloem tissues, corresponding to the predictions from the virtual endoploidy map.

a

#	Marker	Cell type	Predicted 4C ploidy zone Cell # (start-end)	GFP fluorescence boundary position Cell #(mean ± std)
1	SMR1	Atrichoblast	18-25	19±1
		Cortex	22-32	29±1
2	CCS52A1	Trichoblast	1-18	17±1
		Cortex	22-32	23±1

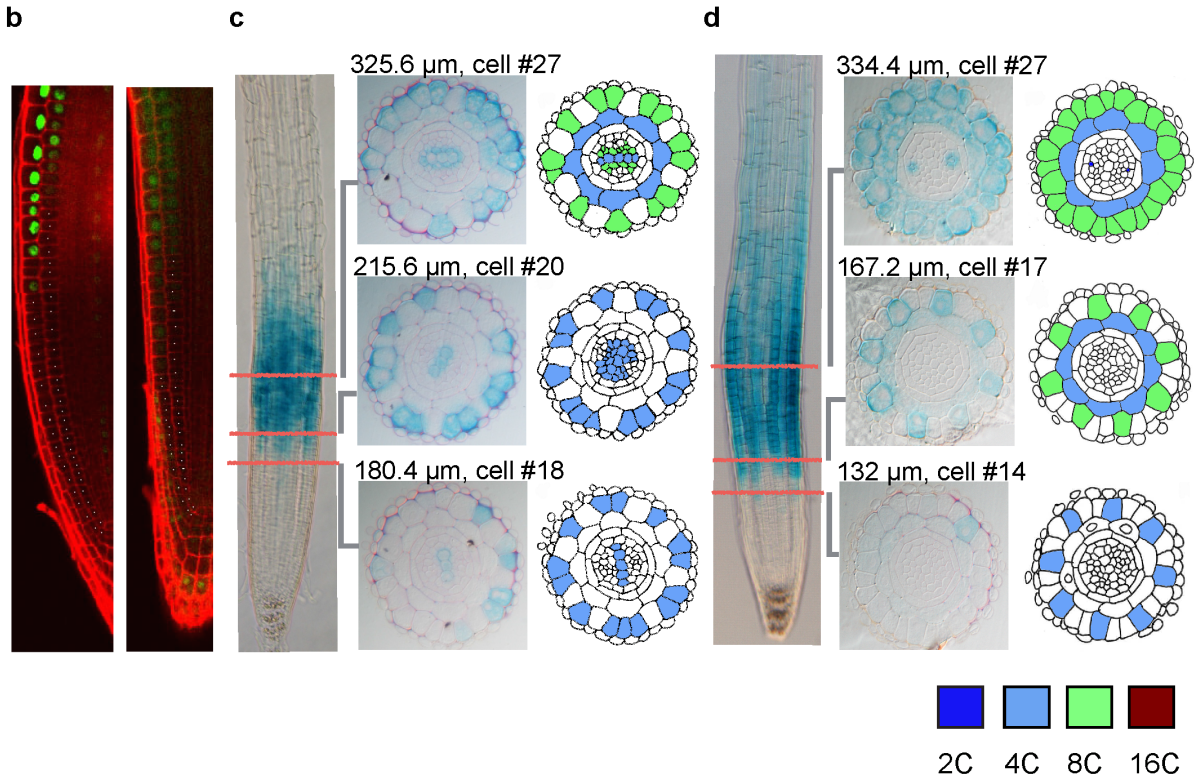


Figure 4.9: Validation for endoploidy borders predicted on the map. (a.) Experimentally determined endoreplication onset boundaries i.e. 4C region positions mapped by measuring the number of cells separating the QC with the first visible GFP signal of SMR1:GFP (**b**,left) and CCS51A1:GFP (**b**,right) marker lines. The white dot indicates the non-GFP cells. **(c,d.)** Experimental mapping of endoreplication onset across tissues within SMR1 **(c)** and SIM **(d)** marker lines. The position of the root cross sections are indicated by red lines. Middle: GUS stained cross sections Distance and cell values indicate the position from the QC and number of atrichoblast cells, respectively. Right: Predicted endoploidy levels of the GUS stained tissues.

4.2.4 Predictions for endoreplicative state change in response to perturbations reveal strong dependence of endoploidy levels on stress signals

The extent of endoreplication is controlled by both environmental[184,229,252] and endogenous factors, such as phytohormones[215,258–260]. Accordingly, gene transcripts peaking at different endoploidy levels showed distinct functional enrichment for hormonal and stress responses (Figure 4.10). This suggested that transcripts whose expression is primarily determined by the endoreplicative state of the cell (as opposed

to transcripts that are stress- or hormone-responsive in a endoreplication-independent way) may be useful in predicting the impact of stress and hormonal treatments on the endocycle program during root development. To assess the value of this approach, we employed a non-spatiotemporal version of our mathematical model to predict the impact of 233 treatments (Supplemental Data Set 7) on the root endoreplication state (Methods Section 4.4.2, Figure 4.11a). Auxin is predicted to have a negative effect on endoreplication, in agreement with previous reports[215]. Treatments with macronutrients (such as phosphate and sulfate deficiency) and micronutrients (such as iron starvation and elevated levels of boron) is predicted to increase the endoreplication level. Similarly, changes in environmental factors (e.g. pH, temperature, and salinity increases) and DNA damage stress (due to UV-B, radiation and genotoxic treatment) were predicted to stimulate endoreplication. The model predictions were confirmed experimentally for salt, low pH, and auxin treatments (Figure 4.11b).

As salt stress ranked top in the list and represents one of the major abiotic stresses, we analysed its effect on the DNA endoploidy level in spatiotemporal detail. To this end, we applied the model on tissue- and stage-specific salt stress transcriptome datasets[261]. Our model predictions suggest that salinity affects the nuclear DNA content distribution across different cell types in a tissue-specific manner. The cells of the cortex and endodermis are predicted to undergo an additional round of endoreplication, whereas the vascular cells show an inhibition of endoreplication; predictions confirmed experimentally (Figure 4.11c; see Methods). Nuclear DNA content predictions in the cells of different root segments indicate that salt affects endoreplication mainly within the transition zone (Figure 4.12).

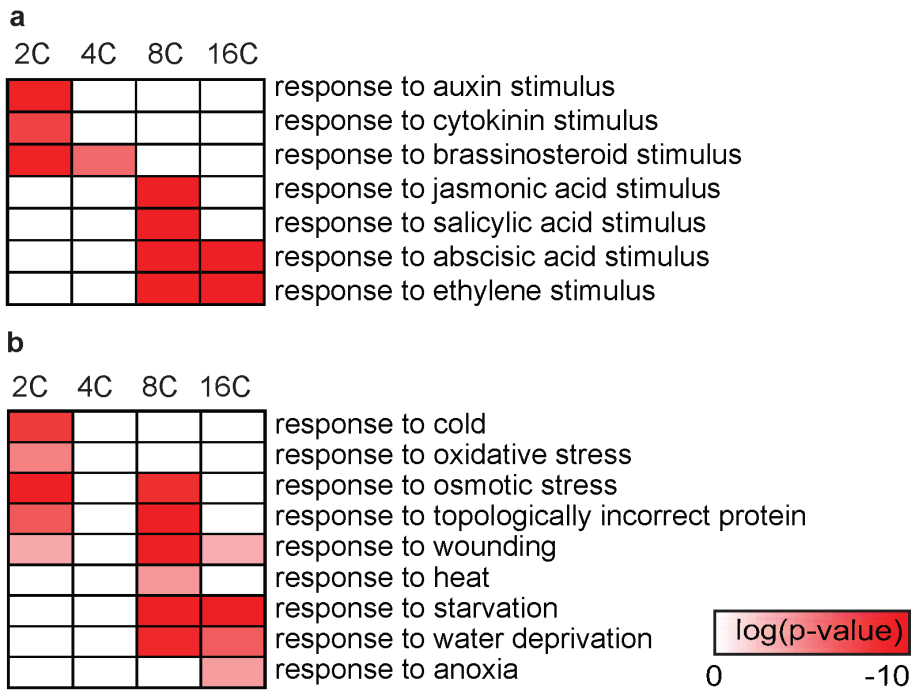


Figure 4.10: Functional enrichment of endoploidy-specific transcripts (i.e. transcripts peak expressed in a particular endoploidy such as 2C, 4C, 8C or 16C) related to (a.) hormone and (b.) stress responses. The differentially expressed 3737 transcripts were grouped based on their peak expression in 2C, 4C, 8C and 16C and studied for their functional enrichment. Here, selected few stress and hormone related GO terms are represented.

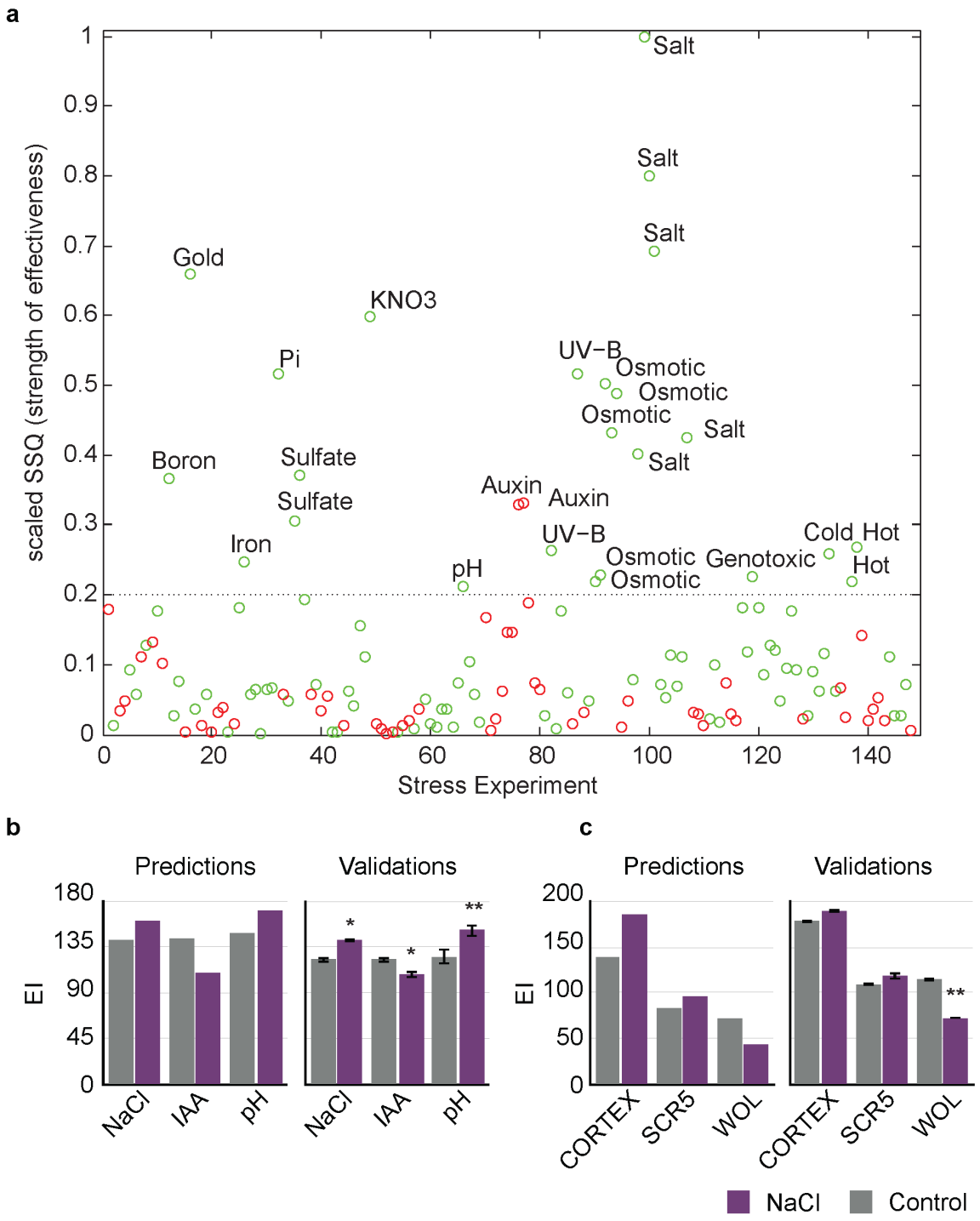


Figure 4.11: The predicted and validated (for representative stresses) effect of stress conditions on change in endoploidy distributions from their respective controls in intact roots and cell types. (a.) The predicted effect of stress conditions on change in endoploidy distributions from their respective controls on intact roots. The x-axis represents 149 publicly available gene expression datasets profiling various stress conditions, and the y-axis provides the endoreplication difference (scaled to 1). Red and green circles indicate that endoreplication is suppressed or promoted, respectively. The dotted line indicates an arbitrary sum of squared errors (SSQ) cutoff (0.2), below which value the conditions were not annotated on the panel. **(b.)** Predicted and validated endoreplication indices (EI) under three representative stresses (140 mM salt, 4.6 pH, and 1 μM auxin (IAA)). **(c.)** Predicted and experimentally validated EI under 140 mM salt and control conditions for cell types (marker) - cortex (COR), endodermis (SCR5) and stele (WOL). The significance of stress effect is judged by two-sample *T*-test. $*$, P value ≤ 0.05; $**$, P value ≤ 0.01.

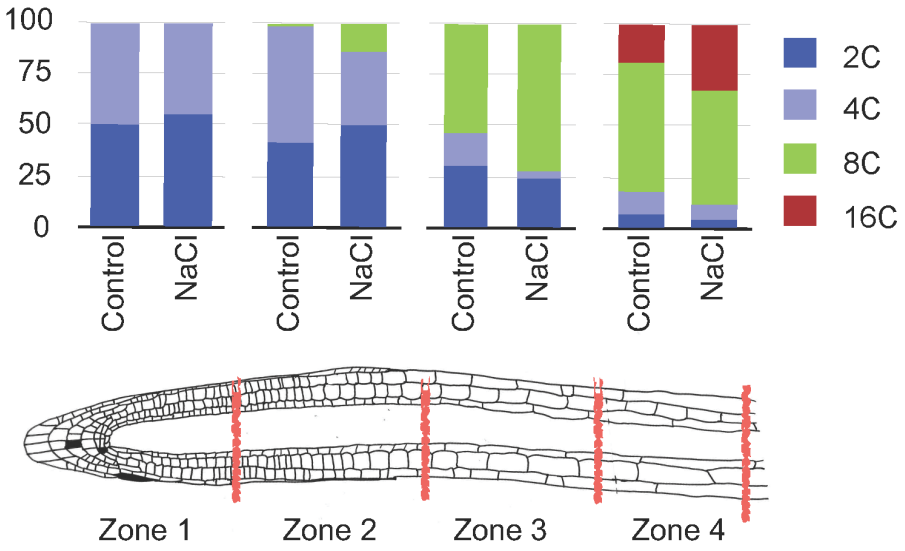


Figure 4.12: Predicted endoploidy distributions for different developmental zones under salt stress (140mM NaCl) and control conditions. The transcriptome dataset measured by Dinneny and co-workers[261] for the developing root cut into four segments i.e. zone-1 (∼150µm from root tip), zone-2 (∼200µm above zone-1), zone-3(∼200-300µm above zone-2) and zone-4 (∼1mm above zone-3) grown under salt vs control condition were used as input to our model II to predict the effect on their endoreplicative state.

4.2.5 Endocycle confers an adaptive response to salinity

The observed changes in nuclear DNA content distribution upon salt treatment suggested that activation of endoreplicative changes might be an integral part of the ability of plants to adapt to salinity. In this context, we studied the salt stress sensitivity of *sim* and *smr1* mutants, in which the endocycle is negatively regulated (Figure 4.13a). The *sim* and *smr1* mutant was highly sensitive to salt compared to Col-0 (Figure 4.13b, c), which suggests that endocycle stimulation upon salt stress represents part of the adaptive response. In future, it would be interesting to further probe the relevance of tissue-dependent endoploidy change, by monitoring the salt stress sensitivity of plants expressing tissue-specifically the CYCA2;3 cyclin, encoding a negative endocycle regulator[157].

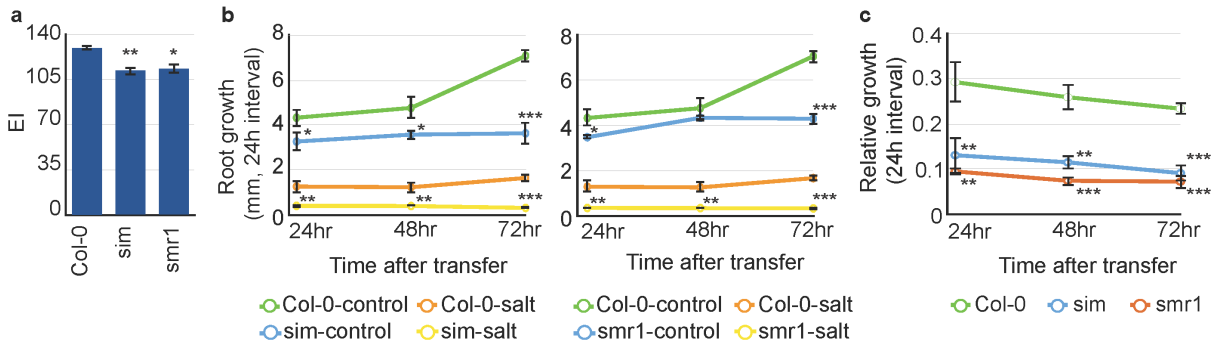


Figure 4.13: Growth measurements of mutant lines under salt and control conditions. (a.) EI, **(b.)** absolute and **(c.)** relative root length of wild type (Col-0), mutant (sim and smr1) lines under salt and control conditions (see methods). The significance of mutation on EI i.e. sim/smr1 versus Col-0 **(a)** and stress effect i.e. sim/smr1 versus Col in control conditions or sim/smr1 versus Col in salt conditions **(b, c)** is judged by two-sample *T*-test. $*$, P value ≤ 0.05; $**$, P value ≤ 0.01; $***$, P value ≤ 0.001.

4.3 Conclusion

Overall, our analysis illustrates a remarkable spatial and temporal control of developmental and environmental cues on endoreplication processes within the root. Additionally, root measurement data on endoreplication mutants grown under salt stress exhibited a role for the endocycle in rapid adaptation to salt stress. Given the correlation between the endoploidy level of a cell and its size[185–187], stress resistance might be linked to physiological differences between small and large cells. Alternatively, growth through endoreplication is speculated to support continuous growth under conditions that limit mitosis[208]. Thus, plant species might use endoreplication to combine the benefit of rapid cell multiplication because of a small genome with the advantage of large cells within specific tissues to withstand stress condition, in which endogenous DNA replication is employed to support an optimal karyoplasmic ratio. The ability of endocycling plants to cope better with stress might explain the evolutionary success of endoploidy, which is mainly observed in annual and biennial species, and in ecological niches that required fast development[240,262].

4.4 Mathematical Models

4.4.1 I : Predicting ST developmental root endoploidy map

Model I simulates the expression patterns of genes in 12 different root slices and 14 different cell types (covered by 17 marker lines) as a function of their endoploidy-specific expression levels in the cortex. The parameters in this model are the endoploidy boundaries along the longitudinal axis of the different root tissues, i.e. the 2C-4C, 4C-8C and 8C-16C boundaries. The boundaries are optimised to give the best possible fit of the modelled gene expression patterns to the measured expression patterns across root tissue marker lines and slices[2]. Our model assumes that, (i) the endoploidy levels (arising through developmentally regulated endoreplication) in each cell type exhibit a logical increment in DNA content over time (2C->4C->8C->16C), (ii) the expression level of a gene in a particular slice is the sum of the gene's expression levels in all cells of different types (and possibly with different endoploidy levels) within that slice and (iii) the measured expression level of a gene in a particular marker line or tissue is the sum of the gene's expression levels in all marked or single tissue cells, respectively, which may have different endoploidy levels. These assumptions are represented in Figure 4.4a. The principal equations of the model are summarised below.

$$E(g,s) = \frac{\sum_{t=1}^{14}\left[\sum_{c\in C(s,t)}\left(\sum_p w_P(t,c,p)\cdot E(g,p)\right)\right]}{\sum_{t=1}^{14}|C(s,t)|} \tag{4.1}$$

$$E(g,m) = \frac{\sum_{t=T(m)}\left\{\sum_{s\in S(m,t)}\left[\sum_{c\in C(s,t)}\left(\sum_p w_P(t,c,p)\cdot E(g,p)\right)\right]\right\}}{\sum_{t\in T(m)}\left(\sum_{s\in S(m,t)}|C(s,t)|\right)} \tag{4.2}$$

In these equations, $E(g,s)$ and $E(g,m)$ represent the simulated expression of gene g in slice s and marker m, respectively. t indexes tissues (cell types). The index p represents the endoploidy level (2, 4, 8 or 16C) and the index c indicates the position of a cell along the longitudinal axis in a particular tissue

(cell type, Table 4.1) t. $C(s,t)$ is the set of cell numbers in slice s and tissue t, as derived from the cell count matrix W_C (see below). $E(g,p)$ represents the endoploidy-specific expression level of gene g at endoploidy p in the cortex dataset. W_P is the endoploidy matrix, with $w_P(t,c,p) = 1$ if cell c in tissue t has endoploidy level p, and $w_P(t,c,p) = 0$ otherwise. The cell count matrix W_C incorporates average cell count estimates obtained from visual inspection of 10 confocal images of *Arabidopsis* wild type (Col-0) roots for the cell types hair cell, cortex and endodermis in the meristematic and elongation zones (slice 1-8), and for the xylem pole pericycle and phloem pole pericycle in the meristematic zone (slice 1-6). The cell counts in the non-hair, phloem, phloem companion, xylem and procambium cell files in slices 1-8 were deduced from the measured counts for other cell types (see Table 4.1 legend for the description). The cell counts in the remaining tissues and slices were based on the cell counts provided in[255]. In equation 4.2, $T(m)$ represents the set of tissues covered in at least some developmental stages (slices) by a particular marker m, $S(m,t)$ being the set of slices in which marker m covers tissue t (Table 4.2).

Table 4.1: The adapted cell counts of 14 distinct cell types in 12 slices (rows 1 to 12). Columns from left to right indicate the cell types quiescent centre (qc), columella (colu), lateral root cap (lrc), trichoblast (hc), atrichoblast (nhc), cortex (cor), endodermis (end), xylem pole pericycle (xpp), phloem pole pericycle (ppp), phloem (p), phloem companion cell (pcc), xylem (x), lateral root primordia (lrp) and procambium (pro).

Slice #	qc	colu	lrc	hc	nhc	cor	end	xpp	ppp	p	pcc	x	lrp	pro
1	4	12	152	24	48	15	12	13	19	0	0	13	0	31
2	0	0	280	37	74	38	39	14	38	14	14	18	0	57
3	0	0	210	32	64	34	38	14	35	14	14	18	0	56
4	0	0	210	28	56	33	35	15	33	15	15	19	0	61
5	0	0	210	25	50	30	31	14	29	14	14	18	0	56
6	0	0	0	21	42	27	29	14	26	14	14	17	0	55
7	0	0	0	54	108	65	83	36	68	36	36	44	0	142
8	0	0	0	16	32	20	29	11	21	11	11	14	0	44
9	0	0	0	40	80	40	40	20	45	20	20	25	0	80
10	0	0	0	40	80	40	40	20	45	20	20	25	0	80
11	4	0	0	40	80	40	40	20	45	20	20	25	130	80
12	0	0	0	40	80	40	40	20	45	20	20	25	0	80

Table 4.2: The marker lines used in mathematical modelling approach and the cell types and cytoplasmic or nuclear tagged GFP marker lines used for validating the *Arabidopsis* root endoploidy map by flow cytometer analysis. For all markers, the slices they cover for a particular cell type are indicated by the range in parentheses. *, indicates marker lines used for flow cytometer analysis. #, indicates marker lines used in this study apart from marker lines described in Cartwright et al. (2009)[255].

Cell-type	Marker-lines	Nuclei/protoplast
QC	AGL42(1), RM1000(1,11), SCR5, *,#WOX5(1,11)	Protoplast
Columella	*PET111(1)	Protoplast
Lateral root cap	*LRC(1-5)	Protoplast
Hair Cell	*COBL9(7-12)	Nuclei
Non-hair cell	*GL2(1-12)	Protoplast
Cortex	J0571(1-12), CORTEX(6-12), *,#CORT(1-12)	Nuclei
Endodermis	J0571(1-12), SCR5(1-12), *,#ENDO(1-12)	Nuclei
Xylem pole pericycle	WOL(1-8), J2661(12), *JO121(8-12)	Nuclei
Phloem pole pericycle	WOL(1-8), J2661(12), *S17(7-12)	Protoplast
Phloem	WOL(1-8), *S32(1-12)	Protoplast
Phloem companion cells	WOL(1-8), *SUC2(9-12)	Protoplast
Xylem	S4(1-6), WOL(1-8), *S18(7-12)	Protoplast
Procambium	*WOL(1-8)	Protoplast

Equations 4.1 and 4.2 essentially sum up the endoploidy-specific expression levels of gene g in all cells in a slice s, respectively marker m, where the endoploidy of each cell (and hence its contribution to the gene's expression level) is determined from the endoploidy matrix W_P. Simulated slice and marker line

expression levels are then compared with the experimentally determined slice and marker line expression levels in[2]. The parameters of the model, the endoploidy boundaries (i.e. the cells at which the endoploidy level changes along the longitudinal axis, as encoded in the endoploidy matrix W_P), are randomly assigned at the beginning of a simulation and optimised using a Monte Carlo-Simulated Annealing (MCSA) strategy to obtain the best possible fit between the simulated and measured[2] expression patterns across all slices and markers (see Section 4.4.3). Overall working scheme of this model is represented in Figure 4.5. In addition, the details of the work flow are represented with an example in Figure 4.14.

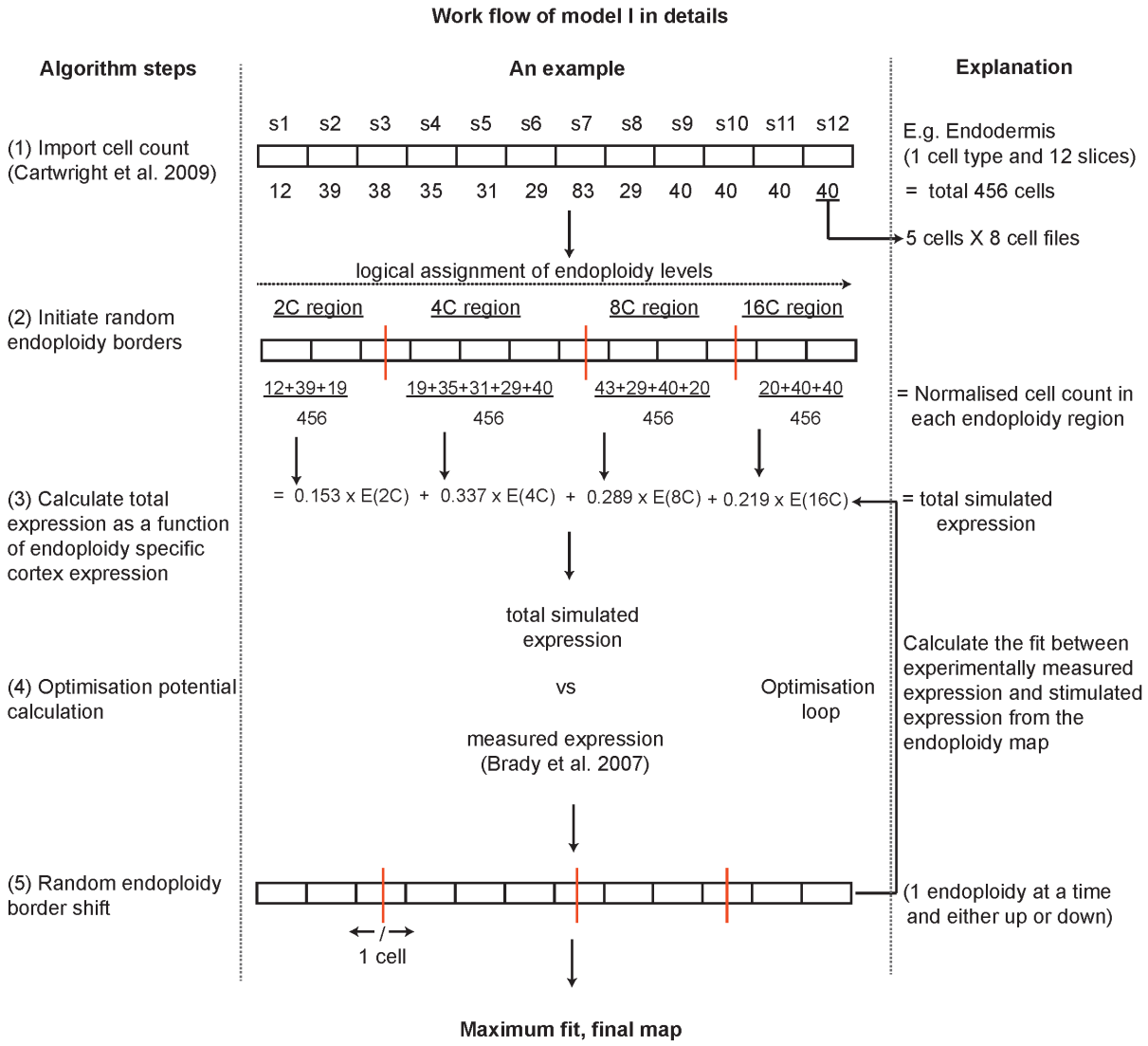


Figure 4.14: Representation of the details of model I work-flow with an example. Here, 1 cell type (Endodermis) and its 12 developmental stages (slices i.e. s1-s12) along longitudinal axis of the root are used as an example to explain the work flow of the model in detail. In the first step, cell count adapted from Cartwright et al.(2009)[255] (Table 4.1) is imported, where cell count in each slice is the total number of cells present in all cell files in that slice. In the next step, model randomly defines the endoploidy boundaries (red lines, 2C-4C, 4C-8C and 8C-16C) by assigning endoploidy levels to each cell in a logical increment (2C->4C->8C->16C). Next, model determines the normalised cell count of each endoploidy region and use it to obtain the total expression of endodermis as a function of endoploidy-specific expression in the cortical cells (E(2C), E(4C), E(8C) and E(16C). Further, the simulated expression is compared with the measured expression by Brady and Co-workers[2] and optimisation potential is obtained. Sequentially, model adjusts one endoploidy border by one cell either up or down along the longitudinal axis of the root to optimise the fit. Finally, model outputs the simulated expression that has maximum fit with the measured expression and the endoploidy map.

4.4.2 II : Endoploidy distribution change prediction upon stress treatment

Model II is essentially a simplified version of model I that simulates the expression patterns of genes in a whole root, a root segment or a particular cell type as a function of the endoploidy-specific expression levels in the cortex dataset. The expression level of a particular gene in any (part of the) root is taken to be a weighted sum of the expression levels in cells with different endoploidy levels in that (part of) the root :

$$E(g) = \sum_{p} w(p).E(g,p) \tag{4.3}$$

where $E(g)$ represents the simulated expression level of gene g, to be compared with the measured expression level in some publicly available microarray dataset on whole roots, segments or cell types subject to a particular treatment, $w(p)$ represents the endoploidy weight i.e., the percentage of cells in the root (segment, cell type) at endoploidy level p, and $E(g,p)$ represents the endoploidy-specific expression level of g at endoploidy p in the cortex dataset. The model uses the same optimisation and simulation strategies as for mathematical model I to optimise the endoploidy weights $w(p)$ (see Section 4.4.3). The details of the work flow are represented with an example in Figure 4.15.

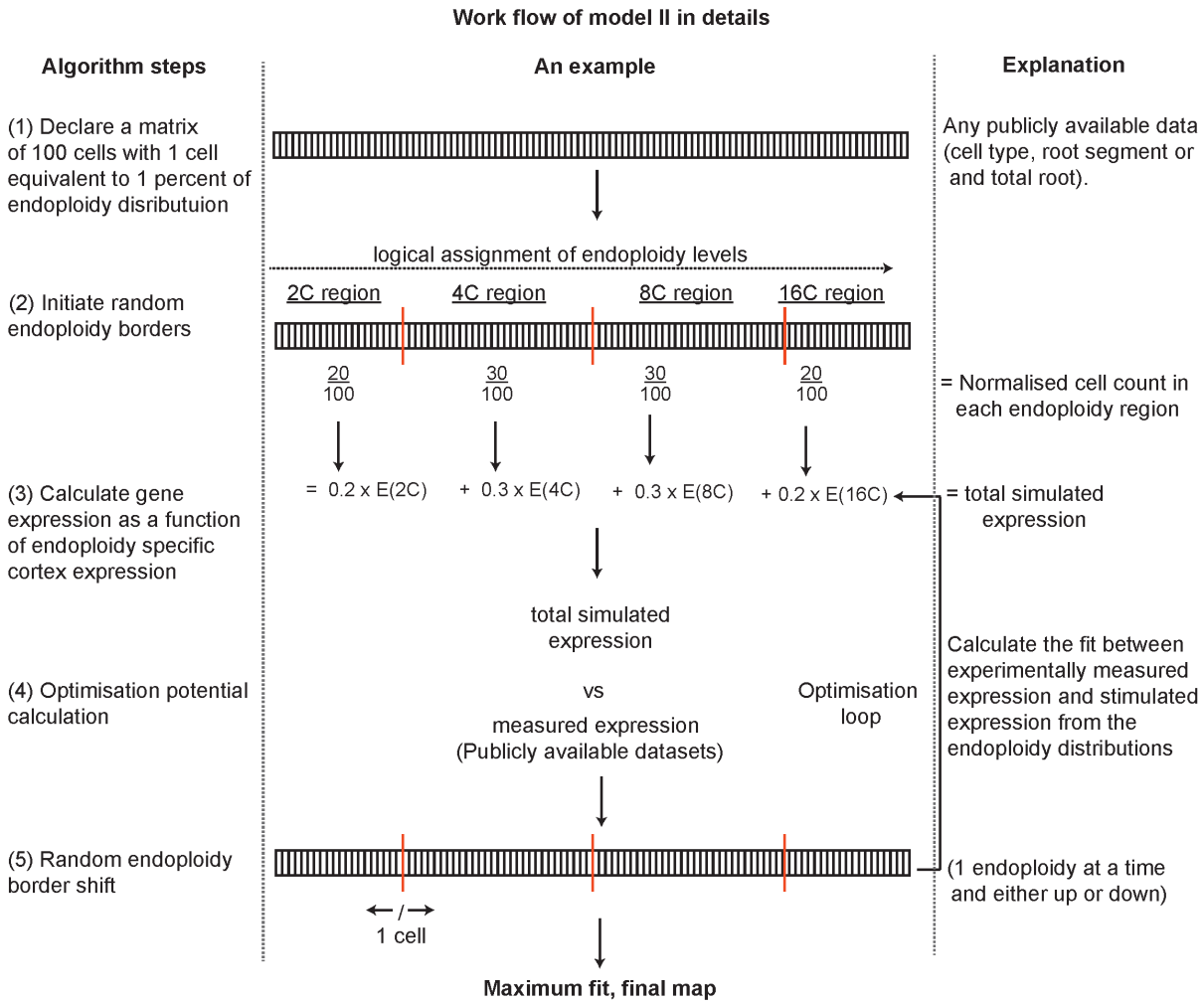


Figure 4.15: Representation of the details of model II work-flow. This model follows the same work flow as described in Figure 4.15, except that it uses a matrix of 100 cells, where 1 cell is equivalent to 1% endoploidy distribution, instead of the actual cell count. This model simulates expression of a particular gene in any profiled cell type, root segment or intact root by optimising the percentage of distributions of individual endoploidy levels.

4.4.3 Simulation and optimisation strategy

We used a classic Monte Carlo Simulated Annealing (MCSA) strategy with exponential temperature decay to optimise the parameters in both Model I and II. The working scheme of the model is represented in Figure 4.5. At the beginning, the parameters, i.e. the endoploidy boundaries in each cell type along the longitudinal axis of the root for Model I or the endoploidy weights for Model II, were randomly assigned, and optimisation progressed by attempting random steps in parameter space, i.e. by moving a particular endoploidy boundary one cell up or down in a single cell type (Model I) or by simulating the effect of a one-cell endoploidy shift on the endoploidy percentages of an entire root, segment or cell type (Model II). A step was accepted if,

$$rand \quad < \quad EXP(-\Delta R/SA) \tag{4.4}$$

with *rand* a random number drawn uniformly from the interval [0,1], ΔR the change in optimisation potential upon taking a step in parameter space, and *SA* the simulated annealing parameter (temperature), which gradually decreases over five orders of magnitude (from *SA* = 10 to *SA* = 0.0001) during the course of an optimisation run, according to the exponential cooling scheme $SA_i = 0.9999\ SA_{i-1}$. The optimisation potential is defined by the reduced chi-squared statistic between measured and simulated expression levels, e.g. in the case of Model I:

$$R \quad = \quad \frac{\sum_{g=1}^{nG}\left[\left(\frac{M_g - simM_g}{\sigma(M_g)}\right)^2 + \left(\frac{S_g - simS_g}{\sigma(S_g)}\right)^2\right]}{[(nG \times 29) - (nP + nG)]} \tag{4.5}$$

where *R* is the reduced chi-squared (goodness-of-fit) statistics, *nG* is the total number of genes *g* included in the simulation, S_g and S_g are the measured expression profiles (vectors) of gene *g* across 12 slices and 17 marker lines, respectively, and $simS_g$ and $simM_g$ are the corresponding simulated expression profiles. $\sigma(S_g)$ and $\sigma(M_g)$ represent standard deviation vectors approximated by element-wise square roots of the S_g and M_g vectors respectively (note that the divisions involving $\sigma(S_g)$ and $\sigma(M_g)$ are also element-wise divisions). The calculated *R* values are divided by a normalisation factor $[nG \times 29 - (nP + nG)]$ i.e. the number of error degrees of freedom. *nP* is the total number of parameters (42 i.e. 3 endoploidy boundaries for each of 14 tissues) and 29 is the sum of 12 slices and 17 markers used.

4.5 Gene set selection

4.5.1 For mathematical model I

In total, 19937 genes are present in the ATH1 transcriptome dataset after preprocessing. However, it is to be expected that not every gene's spatiotemporal expression pattern can be predicted accurately from its endoploidy-specific expression pattern in the cortex, due to e.g. tissue- and developmental stage-specific regulation of gene expression. In other words, not all spatiotemporal gene expression profiles are adequately (i.e. exclusively or to a large enough extent) reflecting endoploidy-specific gene expression changes for the purpose of reconstructing a endoploidy map of the developing root (Figure

4.4b-d). Furthermore, genes that exhibit low endoploidy-specific expression variation (flat profiles in the cortex dataset) are uninformative with regard to detecting endoploidy differences. We therefore took into account several selection criteria for pruning the original set of 19937 genes to a smaller set enriched in genes containing enhanced endoploidy level information. First, genes were selected for high expression levels (> 50th quantile, i.e. >73.17) and high endoploidy-specific expression variation (standard deviation/mean expression >50th quantile, i.e. >20%) in the endoploidy-specific cortex dataset, resulting in a reduced list of 4378 genes. Next, reasoning that reliable endoploidy markers should exhibit endoploidy-determined gene expression levels across many tissues and developmental stages, genes were selected based on whether or not their spatiotemporal expression pattern could be reliably predicted by model I from their endoploidy-specific expression levels, both in terms of quantitative differences (reduced chi-squared statistic, R) and the Pearson correlation (PCC) between simulated and measured[2] expression (Supplemental Data Set 8). For both criteria, we used the 50th quantile (R <179.84 and PCC>0.39) as the selection cutoff, resulting in a further reduced set of 954 genes. Among these genes, 407 peak in 2C in the endoploidy-specific cortex dataset, 162 in 4C, 302 in 8C and 83 in 16C. Since employing unequal numbers of highly expressed genes (markers) for the different endoploidy levels leads to endoploidy-specific biases in the model optimisation runs and the resulting endoploidy map (Figure 4.16), we selected an equal number of genes (83) from each class to obtain a final balanced set of 332 genes to be used in model I (Supplemental Data Set 6).

4.5.2 For mathematical model II

Changes in a gene's expression level in response to stress may be attributed to either changes in endoploidy levels in certain tissues due to stress-dependent modulation of the endocycle (which is our focus here), or to stress responses that are not endocycle-related. To accurately predict changes in the endoploidy distribution upon stress treatments, genes that exhibit stress-responsive expression changes that cannot be attributed to endoploidy changes need to be removed from the aforementioned list of 332 genes used in Model I (see Section 4.4.1). To this end, we first identified publicly available gene expression datasets (stress vs control) (Supplemental Data Set 9) that were generated in a similar experimental setup (except for the stress treatment, but in terms of growth medium, root length etc., to avoid as much as possible expression variations arising due to tissue handling and experimental approaches) as used for the root expression map[2]. Then, we used mathematical model II (see Section 4.4.2) to predict the expression under stress conditions of every one of the 332 genes individually, optimizing the maximum fit to the measured expression profiles under stress given the endoploidy-specific expression levels in the cortex dataset, and we calculated the sum of squared errors (SSQ) between measured and modelled expression:

$$SSQ(g) \quad = \quad \sum_{i=1}^{N} \frac{\left(E_{pred}\left(g,i\right) - E_{obs}\left(g,i\right)\right)^2}{N} \tag{4.6}$$

where, $E_{pred}(g,i)$ and $E_{obs}(g,i)$ represent the modelled and measured expression of gene g at stress condition i, respectively. N represents the total number of conditions used. The SSQ were normalized to the maximum SSQ value over all genes (Figure 4.17). We removed the 9 genes whose stress-responsive expression levels were worst predicted based on their non-stress endoploidy-specific

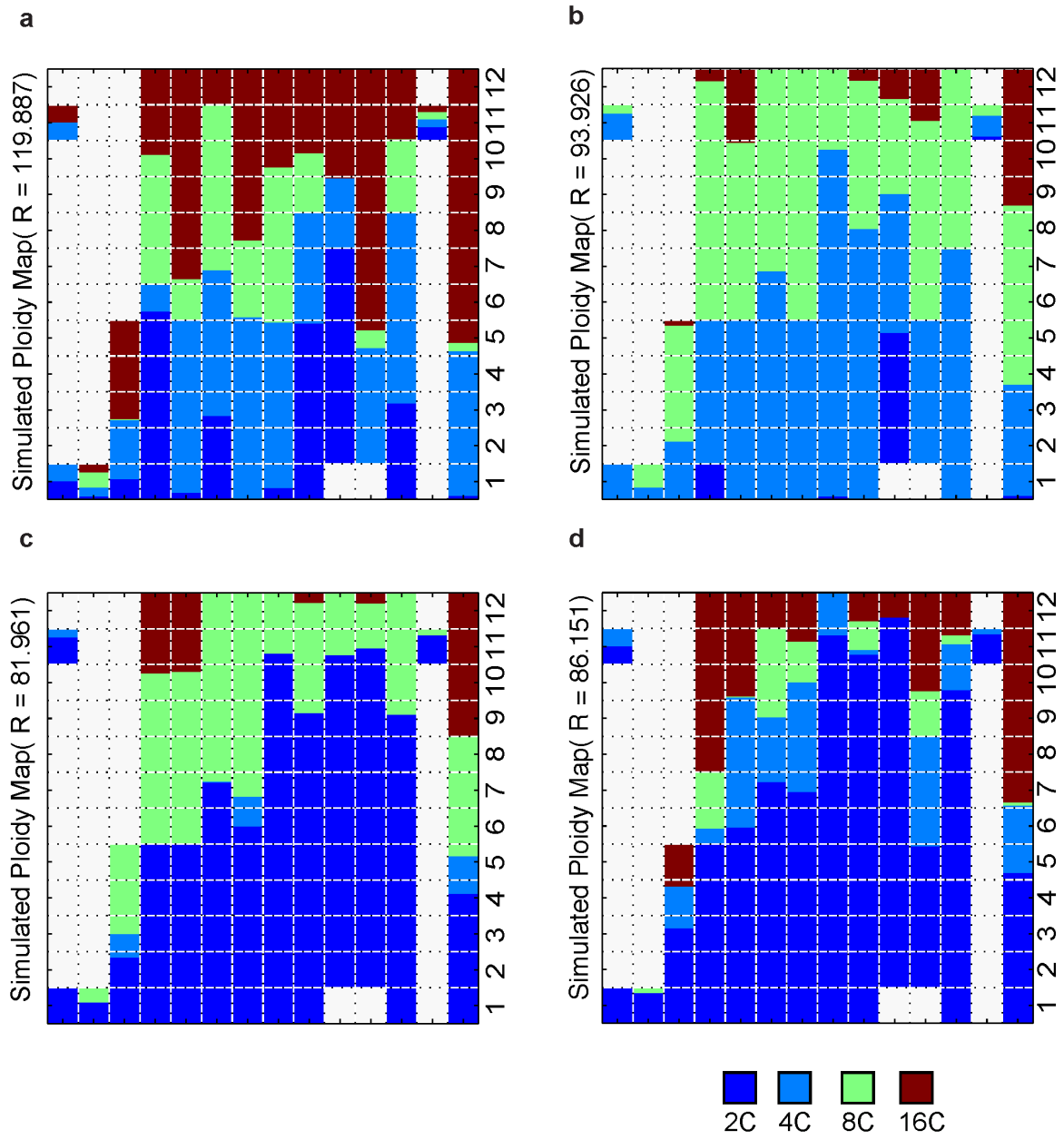


Figure 4.16: Optimized endoploidy maps for selected unbalanced gene sets, i.e. gene sets with unequal numbers of genes (markers) peaking at the various endoploidy levels. Endoploidy maps are shown **(a.)** for the reduced set of 954 genes (see Balanced gene set selection for model I), **(b.)** for a gene set biased towards 2C genes, encompassing 407 genes for 2C and 83 genes each for remaining endoploidy levels, and similarly for gene sets biased towards **(c.)** 4C (162 4C genes, 83 for other endoploidies) and **(d.)** 8C (302 8C genes, 83 for other endoploidies).

cortex expression levels (i.e. the 9 genes with the highest SSQ values) using a visually determined cutoff (Figure 4.17). These genes (highlighted in green, Supplemental Data Set 6) are mainly annotated in the GO database (`www.geneontology.org`, annotation version 27/08/2013) to response to jasmonic acid stimulus, wounding, and salt stress. The remaining set of 323 genes was further used for predicting endoploidy distributions in intact roots (Supplemental Data Set 10), cell types (Supplemental Data Set 11) and root segments (Supplemental Data Set 12) under various stress conditions.

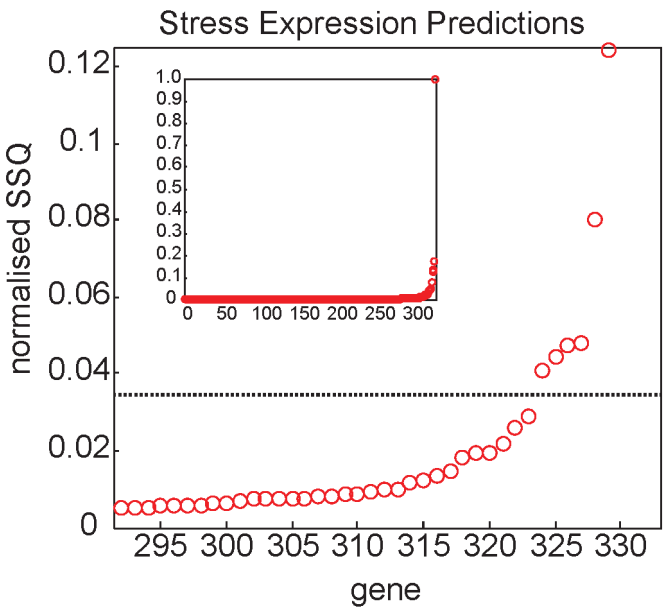


Figure 4.17: Expression prediction performance of 332 genes in selective stress conditions. The Y-axis represents the sum of squared errors (SSQ, normalized to values between 0 and 1) calculated over a set of selected conditions for each gene on the X-axis. The dotted line represents the cutoff used to remove 9 genes which show relatively bad prediction performance under stress in comparison to the remaining 323 genes.

4.6 Materials and methods

4.6.1 Plant lines and growth conditions

Endoreplication-specific gene expression profiles were obtained from sorted nuclei of the cortical cells of an *Arabidopsis thaliana pCO2:YFP-H2b* line[254] (see Section 4.6.3). Tissue-specific endoreplication measurements were obtained via flow cytometric analysis of the thirteen Arabidopsis thaliana marker lines listed in Table 4.2) (previously described in[2,255]). Endoreplication boundary positions were confirmed using reporter lines *CCS52A1*, *SIM* and *SMR1*[213,257]. Root growth measurements under salt and control conditions were performed using *sim*, *smr1* mutant lines.

Seeds were surface-sterilised using a solution of 20 parts by volume of commercial bleach and 80 parts by volume of 100% ethanol, and then washed twice with 100% ethanol. The dried seeds were germinated vertically on plates containing half-strength Murashige and Skoog (MS) medium[263], 1% sucrose and 0.5g/l MES (pH 5.7) in 1% agar. Plants were grown under long day conditions (16h light, 8h darkness) at 22°C. For pH, salt and IAA treatment experiments, plants were grown on a layer of nylon strip embedded on the agar surface to facilitate transfer onto the treatment media. Low pH (4.6), high salt (140 mM NaCl), auxin (1 μM IAA) and respective MS standard media were prepared as described in[261,264,265], respectively. For root growth measurements, five-day-old plants were transferred to 140 mM NaCl and MS standard media and their root lengths were monitored at 24, 48 and 72 hour intervals after transfer.

4.6.2 Flow cytometer analysis

Sections from five-day-old roots were excised with a razor blade approximately 0.5 cm below the root tip. The excised tips of cytoplasmic lines (Table 4.2) were incubated in 8 ml of protoplasting solution [1.25% Cellulase (Yakult, Japan), 0.3% Macerozyme (Yakult, Japan), 0.4 M mannitol, 20 mM MES, 20 mM KCl, pH 5.7 adjusted with 1 M Tris/HCl pH 7.5, activated at 55°C for 10 min and then cooled to room temperature, 0.1% bovine serum albumin and 10 mM CaCl2] in 25ml Erlenmeyer flasks for 2 hours on an orbital shaker (100 rpm) under continuous light. The protoplasts were then filtered through a 40 μm filter and centrifuged at 1000 rpm at 4°C for 10 min. The pellets were resuspended in 1 ml of wash buffer (identical composition to that of the protoplasting solution but lacking the enzymes and activation pretreatment). The GFP-expressing protoplasts were FACS sorted and collected in 200 μL CyStain UV Precise nuclei extraction buffer (Partec) and their nuclei were stained by adding 800 μL nuclei staining buffer (Partec). DNA content of GFP-expressing protoplasts were measured with a CyFlow Flow Cytometer (Partec) exited by illumination at nm and analysed with the FloMax software (Partec).

The cut root tips of the nuclear lines (Table 4.2) were further chopped with a razor blade in 200 μL of nuclei extraction buffer containing 45 mM MgCl2, 30 mM sodium citrate, and 20 mM 3-morpholinopropane-1-sulfonic acid, pH 7.0[142] for 2 mins, then filtered through a 50 μm nylon filter. The DNA was stained with 1 mg/ml DAPI (4',6-diamidino-2-phenylindole[266]). Nuclei were measured using a CyFlow Flow Cytometer (Partec) exited by illumination at 395 nm, and equipped with an additional 488 nm laser to excite and detect GFP-specific fluorescence. The measured DNA contents were analysed using FloMax software (Partec).

4.6.3 Endoploidy-specific Microarray data acquisition

Five-day old roots of *pCO2:YFP-H2B* plants (Col-0 ecotype) grown under continuous light conditions at 22-23°C were excised using a razor blade at approximately 3/4 from the root tip. Samples of combined root material (10g fresh weight) were collected in a glass petri dish and chopped with slicing action after adding 10 ml nuclear isolation buffer (45 mM MgCl2, 30 mM Sodium Citrate (trisodium), 20 mM MOPS (3-LN-morpholino propanesul fonate), adjust pH to 7.0). The root material was then transferred on a 100 μm strainer in a 50 ml tube (the petri dish was rinsed with nuclear isolation buffer to yield 15 ml volume). The solution was collected into a 15 ml tube and centrifuged at 2500 rpm in a Sorvall swinging bucket AH-3000 at 4C for 8 min. The pellet was drained to about 0.5 ml and resuspended in nuclear isolation buffer to 4 ml, then transferred to a 30 μm strainer in a 5 ml tube and DAPI was added (20 μl/ml 0.1 mg/ml stock). Biparametric sorting was then done based on YFP fluorescence versus nuclear DNA content, as previously described by Zhang and co-workers[266]. Sorting of isolated nuclei was done using a Dako-Cytomation MoFlo flow cytometer/cell sorter as described by Zhang and co-workers[267]. The nuclear RNA was extracted from each nuclear DNA content population (2C, 4C, 8C, and 16C, approximately 0.2 ml [200,000 nuclei]/0.95 ml RLT) using Qiagen RNEasy kits according to the manufacturer's instructions. Prior to Affymetrix ATH1 array hybridisation, two consecutive rounds of RNA amplification were done in the MAF (VIB Microarray Facility, http://www.nucleomics.be), using standard Affymetrix protocols for small samples. The amplified nuclear RNA of each DNA content class was used for microarray analysis.

4.6.4 Normalisation and Data analysis

The raw endoploidy-specific microarray data was preprocessed using the Robust Multichip Average (RMA) normalisation approach (background correction, quantile normalisation and summarisation) implemented in the Bioconductor R package, version 2.5[29,35]. The Bioconductor package limma[34] was used to identify differentially expressed genes. Pairwise comparisons between any two endoploidy levels were performed using moderated *t*-statistics and the eBayes method as implemented in limma. P values were corrected for multiple testing using the Benjamini-Hochberg method[268] at a false discovery rate (FDR) threshold of 0.05. The differentially expressed genes were k-mean clustered using the Matlab function 'kmeans' with 'correlation' as the distance measure, 24 (i.e. the total number of possible expression level rank patterns over the four endoploidy levels) as the number of clusters to be obtained and 50 repeats. The centroid pattern of each of these clusters is represented in Figure 4.1. These clusters were further classified as endoploidy-specific based on their peak expression. Functional enrichment was calculated with the BiNGO tool[40] using hypergeometric tests and Benjamini-Hochberg correction at FDR = 0.05. The ST peak expression of any cluster was profiled based on root cell type- (marker-) and developmental stage-specific gene expression datasets[2] (Figure 4.2).

The same ST expression datasets were used in addition to the endoploidy-specific cortex dataset in mathematical model I (see Section 4.4.1). To this end, the raw data from all three datasets (marker-, slice- and endoploidy-specific) were jointly RMA normalised as described above. The untransformed (i.e. non-log-scale) expression values were used in the model. We considered 17 markers covering 14 tissue types and 12 slices as reported in Cartwright et al. (2009)[255]. For mathematical model II (see Section 4.4.2), raw microarray datasets for stress experiments on whole roots, root segments and particular root cell types were obtained from CORNET[131,269] and the GEO repository[270] (Supplemental Data Set 7). Each of these raw datasets was RMA normalized separately with the endoploidy-specific cortex data as described above. Again, untransformed expression values were used in the mathematical model.

4.6.5 Endoploidy map validation experiments

Flow cytometer experiments

The endoploidy content of cells of different tissue types in 0.5 cm-long *Arabidopsis* root tips (Figure 4.7) was measured using flow cytometer analysis (see Section 4.6.2) on cell type-specific GFP marker lines. As the expression profiles of most cell types were covered by two or multiple marker lines[255], we used marker lines that cover the later developmental stages (Table 4.2). We used the measured endoploidy profiles to locate the endoploidy boundaries in the tissue and developmental stages covered by a particular marker line, and inferred the endoploidy levels for earlier developmental stages (Figure 4.8). The cell count (represented in Table 4.1) of a particular cell type in a particular developmental stage is a sum of cells from all cell-files of that cell-type in that slice. In the model as well as for validation experiments, we assumed that all cell-files of a particular cell-type along longitudinal axis do not undergo endoreplication at the same time (at the same cell number). To represent the measured endoploidy distributions in a manner similar to the predicted map (Figure 4.6), the total number of cells (i.e. sum over 12 developmental stages) in a particular cell type are treated as a pool of cells and distributed

in the four endoploidy classes using measured endoploidy distributions. These cells are then logically (2C->4C->8C->16C) arranged in a temporal manner i.e. re-distributed over 12 slices using the cell count matrix provided in Table 4.1. The endoploidy boundary positions (in both predicted and validated map) in a particular cell-type are obtained by dividing the total number of cells (over 12 developmental stages) by the total number of cell-files in that particular cell-type.

Endoploidy border analysis

The endoploidy boundary positions observed in the validated map are not always reliable, as they are highly dependent on the absolute number of nuclei or protoplasts extracted and measured per root in the flow cytometer analysis. The efficiency of protoplasting and nuclear extraction are dependent on the duration with which tissues are treated with the respective extraction buffers. Such dependency often gives quantitative variations in the endoploidy distributions, and it is not straightforward to compare data obtained from different experiments. In addition, the cells in the maturation zone are tightly packed compared to those in the meristematic zone, and might therefore be only partially extracted in a given timeframe, leading to the further measurement errors. Moreover, the flow cytometer cannot distinguish the mitotically dividing G2 cells from the 4C cells, leading to uncertainties in the location of the 2C-4C boundary in any given cell type. To overcome these technical limitations, we used confocal microscopy to locate the first cell from QC exhibiting a visible GFP signal of two endocycle markers *SMR1* and *CCS52A1* and several endoploidy-specific markers (either 4C, 8C or 16C) tagged with GFP. We performed this analysis on 3-5 roots for each marker line and the average first cell number was taken as the endoploidy boundary cell number estimate (Figure 4.9a).

Endoreplication onset order analysis

To validate the order of endoreplication onset among 14 different cell types, we studied root cross-sections at various distances from the QC in GUS stained endocycle marker lines (*SMR1* and *SIM*). First, we generated a cell distance matrix of a cell types atrichoblast by counting the cells along the longitudinal axis of a root and measuring their distance from the QC. We repeated this analysis for 30 different roots to generate an average distance map (Supplemental Data Set 13). Next, we used the measured average distance matrix to identify the cell numbers of atrichoblast cell type visible in particular root cross-sections. The measured endoreplication onset order was compared with the predicted order mapped on a virtual 2D root (Figure 4.9b,c).

4.7 Acknowledgement

We thank high-performing-cluster (HPC) facility at University of Ghent for providing access to the infrastructure. This research and R.B. was supported by the Fund for Scientific Research-Flanders Grant G.0029.11 to S.M. and L.D.V.. We thank Prof.Dr.Tom Beeckman (VIB, University of Ghent, Belgium) and Prof.Dr.John C. Larkin (Louisiana State University, USA) for providing seeds of tissue-specific marker lines and sim and smr1 mutant lines, respectively, used in this study.

4.8 Author contributions

L.D.V. and S.M. conceived the research. R.B., V.B., S.M., and L.D.V. designed the work. R.B. performed and S.M. supervised the computational work. R.B., V.B., I.V., and F.C. performed and L.D.V supervised the experimental work. G.V.I., V.B., G.M.L., and D.W.G. generated the transcriptome data. R.B., S.M., and L.D.V. analysed the data and wrote the manuscript.

Chapter 5

Summary, future perspectives and applications

“Imagination is more important than knowledge”

Albert Einstein.

For the author contributions, see page 106.

5.1 Summary

Predictive modeling approaches allow us to formulate testable hypotheses from multidimensional data inputs, which in turn can be experimentally verified. In this thesis, we showed two such approaches that exploit microarray transcriptome data to answer questions that were difficult to address experimentally.

The first approach predicts gene functions from subtle uncontrolled expression variation among individual wild-type *Arabidopsis* plants. The standard gene function prediction approaches use gene expression data generated from traditional profiling experiments, in which plants that are grown under tightly controlled experimental conditions subjected to harsh treatments and pooled. However, in natural context such experimental setups are unrealistic as individual plants are simultaneously exposed to subtle changes in multiple environmental conditions. Thus, as an alternative to traditional experimental setup we studied the expression variations due to subtle uncontrolled perturbations among individual wild-type *Arabidopsis thaliana* plants grown under the same macroscopic growth conditions. We found that the underlying gene network structure of expression profiles from an alternative setup contains as much information as the traditional setup. We also found that subtle uncontrolled variations in gene expression between individuals could be used to predict functional links between genes and unravel regulatory influences, which could lead to the implementation of candidate gene identification strategies with lower effort and costs than traditional gene expression profiling setups. We finally showed the use of this approach to identify and validate *ILL6* as a new regulatory component of the jasmonate response pathway.

The second approach uses gene expression data to predict a ST DNA endoploidy map of the *Arabidopsis* root. Plant organ development involves tight co-ordination between cell cycle and endocycle. During endocycle, cells skip mitosis, resulting into increased endoploidy and cell size. This process is essential for cell fate maintenance and response to physiological conditions. However, an adequate knowledge on the arrangement and functional relevance of dividing versus polyploidal cells in a developing organ was missing in our understanding of the process. In this context, we developed a mathematical model to predict an ST endoploidy map of the *Arabidopsis* root at cellular resolution and further validated it using cell biology experiments. The map reveals that the order of onset and the extent of endoreplication is distinct among cell types at different developmental stages. For example, the outer cell types undergo endoreplication earlier and more extensively than the inner ones. Further, we demonstrated that stresses such as salt stress affect the arrangement of endoploidy regions in the root during the adaptive growth response. Overall, we established that the endocycle is tightly regulated and contributes to root growth in response to both developmental and environmental cues.

5.2 Future perspectives and applications

5.2.1 A data sampling approach to assign functions to candidate genes

There are two main approaches usually used to assign functions to candidate genes based on expression data. A first approach is to examine a large compendium of expression profiles including all possible conditions and extract gene function predictions from it using network-based or module-based algorithms[88]. An important disadvantage is that informative experiments in a particular context (i.e. for a particular biological process) may be non-informative in another context and use of amalgamated datasets may dilute the positive effects of informative experiments on the predictive power. In the second approach, function predictions for unknown genes are done by analysing experiments probing a specific set of conditions related to the process of interest. Although this approach is more focused on the process of interest, chances are that some relevant genes are not predicted as candidates because links between these genes known to be involved in the process of interest appear only in other than the included conditions. In other words, this technique ignores the fact that relevant information on a particular process may surface under conditions that at first sight have nothing to do with the process of interest. The aforementioned conflicting problems in gene function prediction approaches, may be circumvented by the novel experiment sampling technique used in Chapter 2. This technique randomly samples conditions from a pool of conditions, thereby resolving the data dilution problem (in at least some samples) without sacrificing the extent of data analysed over all sampling repeats.

Figure 5.1 shows preliminary results on the global function prediction performance (obtained as described in Chapter 2) of 1000 networks obtained from randomly assembled compendia of sample size of 50, 70 and 100 experiments (from a pool of 1044 experiments) compared with (i) networks targeted for compendia for response to bacteria and temperature stimuli and (ii) networks obtained from large compendia probing expression responses in biotic environments (BE), abiotic environments (AE) and various plant growth hormone regimens (PGHR), and a mother network (BIG) that includes all of the 1044 experiments retrieved from the Cornet database[131,269]. The results indicate that a random samples of sufficient size can on average outperform a targeted compendium (for all FDR thresholds 10E-2 to 10E-11) in terms of functional prediction confidence (here a sample of 100 experiments seems to be sufficient). Sampled networks of size 100 seem to have better performance on average than that of the BE and PGHR compendia for all FDR thresholds. For the BIG and AE compendia this holds true only for the higher (or less stringent) FDR thresholds (10E-2 and 10E-3), but even for the more stringent FDR thresholds, a few of the sampled compendia networks are able to achieve a similar performance score (F-measure) to that of BIG and AE networks. This result suggest some smaller samples can always substitute for large compendia without losing much prediction power. Similar results are seen for specific GO categories such as response to water deprivation (Figure 5.2). In this particular case, the average samples are outperforming the BIG compendium.

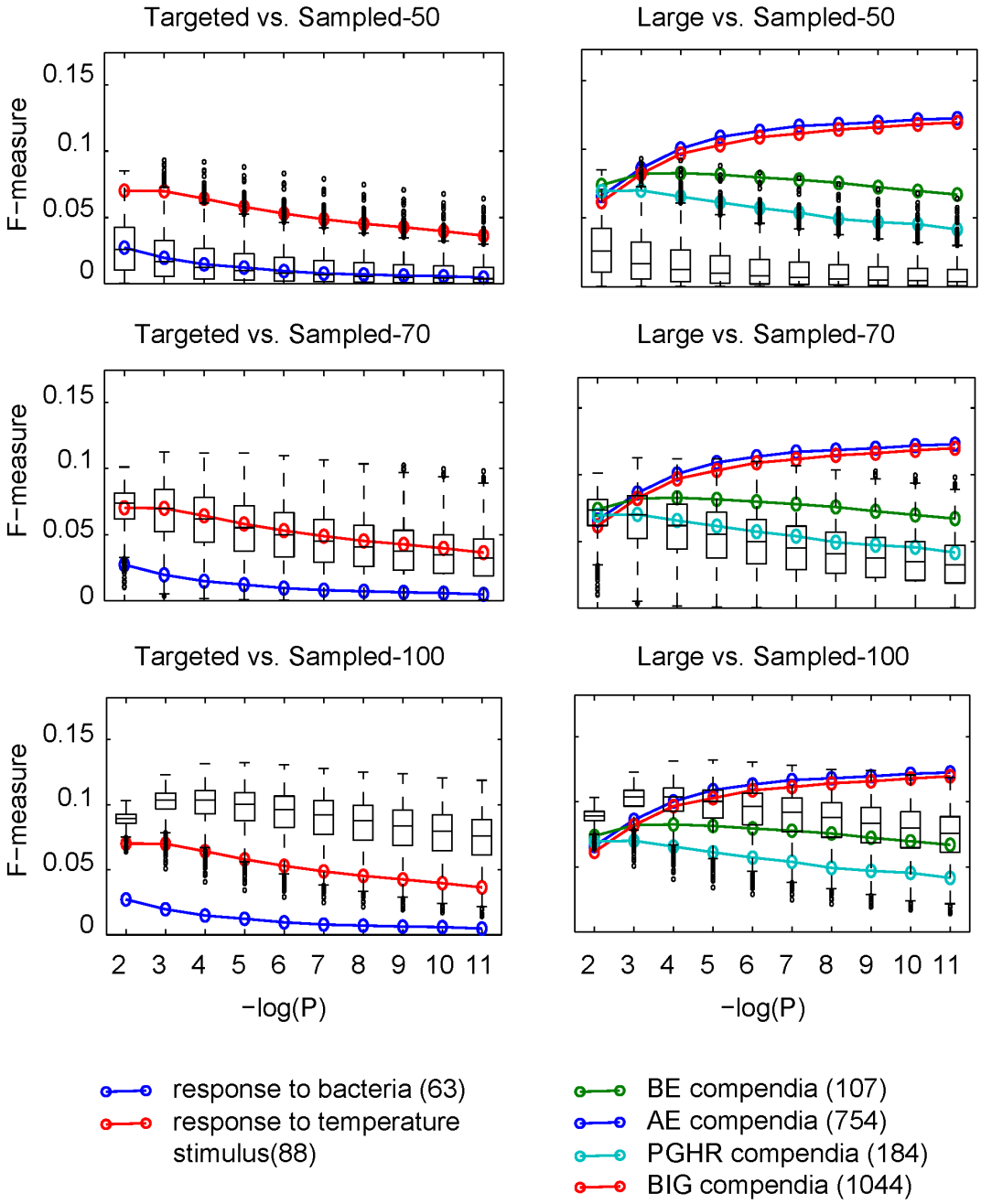


Figure 5.1: Global function prediction performance comparison between sampled networks and targeted and large networks. The performance of the sampled networks of size 50, 70 and 100 (box-and-whisker plots, see legends) is compared with the targeted i.e. bacteria and temperature regimen networks (left column) and the networks retrieved from large compendia, i.e. BE, AE, PGH and BIG (right column, open circles and solid line, see legends). The global performance of a network is measured as described in chapter 2. In the legend, the total number of experiments for each compendia are specified in parentheses. For description of x- and y-axis labels see legend of Figure 2.8.

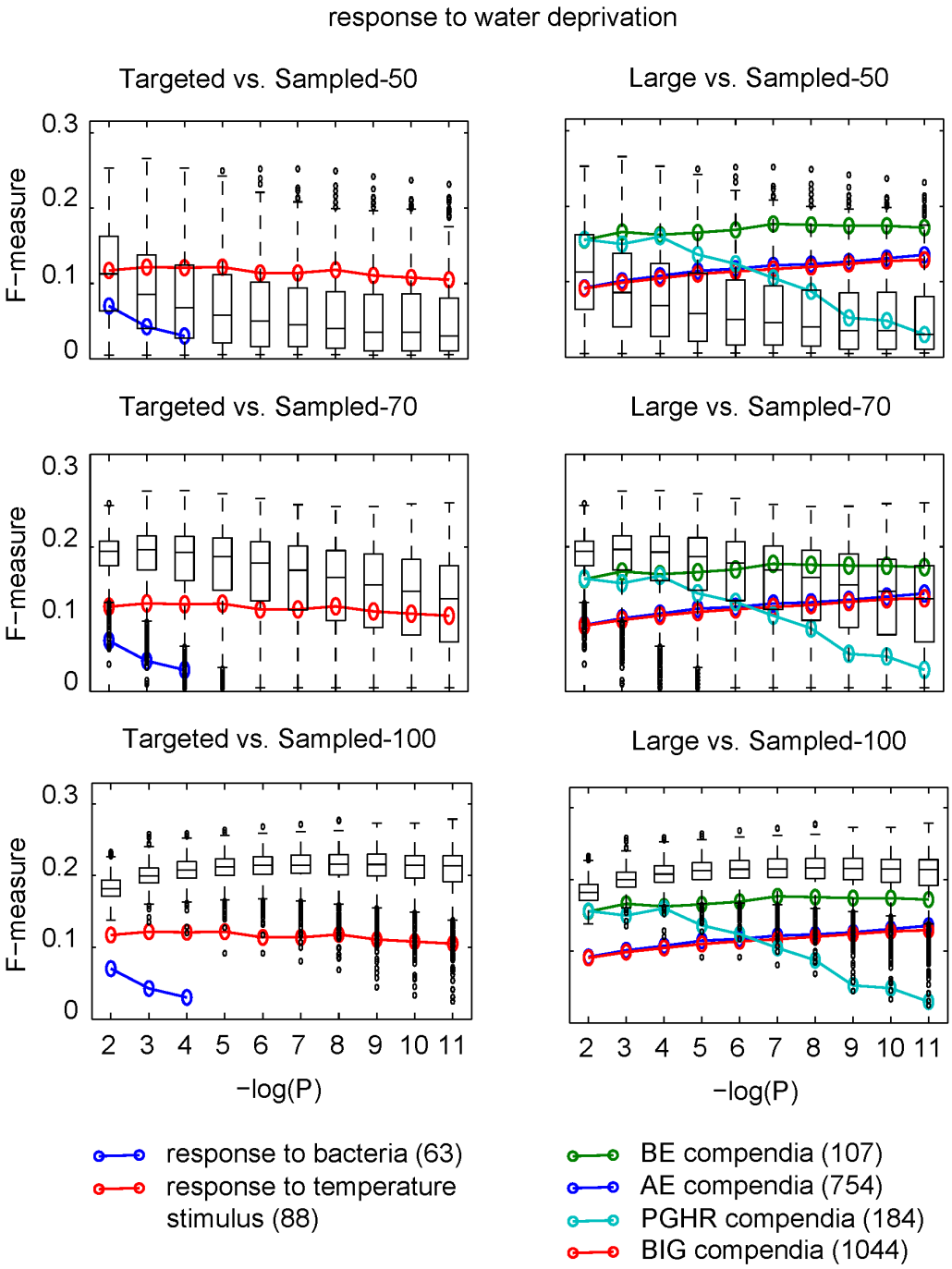


Figure 5.2: Function prediction performance for GO category response to water deprivation. The performance of the GO category "response to water deprivation" of the sampled networks of size 50, 70 and 100 (box-and-whisker plots, see legends) is compared with the targeted i.e. bacteria and temperature regimen networks (left column) and the networks retrieved from large compendia, i.e. BE, AE, PGH and BIG (right column, open circles and solid line, see legends). The performance of an individual GO category is measured as described in Chapter 2. In legend, the total number of experiments for each compendium is specified in parentheses. For description of x- and y-axis labels see legend of Figure 2.8.

Potential work

In order to compare the performance of sampling approach vs large and targeted compendia approach, the information from all 1000 sampled compendia needs to aggregated in one ranking of prioritised gene function predictions. Therefore, different methods for prioritisation, for e.g. as based on optimal P values or on number of times a gene is predicted across samples, can be investigated. The ranked lists of predictions from sampled, large and targeted compendia can be compared for the true positive and false positive rates and f-measures (obtained by comparing predictions with reference GO) to assess the prediction performance of sampling approach. Additionally, individual sampled compendia that are outperforming both targeted and large compendia can be investigated for the kind of experiments which they contain, and whether these are the usual suspects anticipated to harbour information on the target process or not. This exercise will aid wet-lab scientists in setting up appropriate experiments for discovery of genes involved in their process of interest.

Broader significance

As understanding of plant growth regulation is a basis for developing improved crop varieties for stress tolerance and yield, a better and more systematic understanding of the complex regulatory networks that govern their development and environmental plasticity is necessary. The reported work in Chapter 2 and successive perspectives described here might be useful to identify new regulatory components as well as assign functions to known components involved in various biological processes. Additionally, the sampling approach described above might suggest novel experimental setups (specific to the process of interest) to uncover functional links between regulatory components and regulatory influences. The techniques developed here could greatly boost our understanding of the wiring of plant systems involved in stress responses and yield and may help develop improved plant varieties.

5.2.2 Diagnostic markers to predict endoploidy distribution change in *Arabidopsis* root in response to environmental and endogenous factors

The endocycle is known to be modulated by environmental factors such as sunlight, high temperature, shade and water deficit etc.[184,229,252] and endogenous factors such as the levels of auxin, cytokinin, gibberellic acid and abscisic acid (ABA)[215,258–260] (Figure 3.6). However, the change in ST arrangement of polyploid vs dividing cells in a developing organ in response to stress conditions and its physiological relevances are still poorly understood. In Chapter 4, we successfully investigated such ST endoploidy arrangements in *Arabidopsis* roots under developmental cues, and the impact of stress and hormonal treatments thereon, based on gene expression data of 332 predictor genes. In principle, a small representative set of marker genes can be obtained from these 332 genes by removing one gene at a time, with the aim of predicting same endoploidy map obtained using the original gene set. The smaller size of the new gene set provides advantages such as the ability to predict the endoploidy map of the developing root under various environmental stress conditions, since small sets of transcripts can be measured using cheap, accurate and convenient techniques such as nano-string instead of RNA-Seq or

micro-arrays. In addition, generating multiple alternative but equally performing smaller sets may tell us about the (classes of) genes or markers that are important for the endoreplication process.

Figure 5.3 shows a preliminary approach (Figure 5.3-a & b) used to obtain a small set of 70 predictor genes and the resulting optimised endoploidy map (Figure 5.3-c). The Euclidean distance between the two optimised maps learned using the original gene set (332 genes) and the smaller set of 70 predictor genes suggests that both maps are very similar (Figure 5.3-d). This analysis suggests that small diagnostic marker sets can be reliably used to predict endoploidy maps at a ST level. Figure 5.4 shows the comparison between predicted endoploidy distributions using the original set (323 genes, Figure 5.4-a,b & c left column) and the small diagnostic marker set (70 genes, Figure 5.4-a,b & c right column). Although there are quantitative differences among the predicted endoploidy profiles, they qualitatively show the same trend, namely that salt treatment differentially regulates endoreplication in a cell type and developmental zone specific manner and overall increases the total endoploidy content of the intact root. Overall, these preliminary results illustrate the potential of small diagnostic marker sets in predicting endoploidy distributions in intact roots, cell types or segments of the root.

Potential work

Although we identified 332 markers that are important in predicting endoploidy distributions in intact root, cell types or root segments under developmental and environmental cues, the relevance of most of these markers (among 332) in endoreplication process is still missing. In principle, this list of markers can be further narrowed down to a small realistic number of markers for experimental investigation. The predictor set described above was obtained by removing one gene at a time, while optimising for the same endoploidy border positions predicted using the set of 332 genes. Due to the nature of this method, where genes were prioritised for removal based on their optimisation potential (i.e. difference between endoploidy border positions predicted using 332 genes and a particular set of genes after removing a particular gene), multiple repeats would result in relatively similar predictor sets. Other methods with more randomised approaches for pruning marker gene sets can be investigated for obtaining more diverse small marker sets, which may reveal whether particular markers (presumably endoreplication process specific) are preferentially retained over multiple marker sets. These markers would be the ideal candidates for testing experimentally (using mutant lines) their relevance in the endoreplication process. Small sets can be further used to predict endoploidy maps under control and stress conditions at ST level using transcript data generated from nano-string technology. In this context, we plan to measure transcripts in four cell types (trichoblast, atrichoblast, cortex and endodermis) and three developmental stages (meristematic, elongation and differentiation) under control versus stress conditions such as pH, temperature, genotoxic, etc. The normalised and processed data will be used as input for the non-spatiotemporal version of mathematical model described in Chapter 4 for making predictions, and which will be further validated using flow cytometry experiments. On the modelling front, the spatio-temporal version of our mathematical model (described in Chapter 4) can be refined by incorporating information regarding the nuclear-cytoplasmic ratio i.e. endoploidy by cell volume. However, measuring real cytoplasmic volume (apart from vacuoles) could be a daunting task.

Broader significance

Roots provide the basis for survival of the whole plant as they structurally support the aerial portions, acquire nutrients and water essential for plant growth, synthesise hormones, and are the site of interaction with soil bacteria. Thus, understanding the regulation of root growth and development is an important task in plant biology. The fundamental work described in Chapter 4 and successive future perspectives described here will improve our understanding of the effect of stress conditions on root development at cellular resolution, which will underpin the efforts in identifying important traits of root development and in developing better crop varieties.

5.2.3 Predicting endoploidy distribution patterns during *Arabidopsis* leaf development

Leaves are an important part of the plant as they play a pivotal role in photosynthesis, respiration and photo-perception. Their growth involves three distinct phases: leaf primordia development, primary and secondary morphogenesis. During primary morphogenesis, growth is sustained by successive cell divisions, and subsequently by cell expansion in secondary morphogenesis. The transition from cell proliferation to expansion is often marked by endocycle onset. Thus, the balance between cell proliferation and cell expansion determines the total number of cells and the final leaf size. Earlier it was believed that the transition from cell proliferation to expansion proceeds in a gradient down the leaf, with cell proliferation first ceasing at the tip and then progressively down the longitudinal axis[192]. Contrary, recent detailed kinematic and transcriptome analysis of six developmental stages illustrated that this transition occurs abruptly[193]. Nevertheless, a clear view on the extent of endoreplication during these developmental stages is still missing.

In this context, the available transcriptome data of leaf developmental stages can be used to reverse engineer the distributions of proliferating and endoreplicating cells as per the modeling approach used in Chapter 4. Prof. Lieven De Veylder in collaboration with Dr. Katja Bärenfaller (ETH, Zurich) have recently obtained endoploidy-specific transcriptome data as well as total expression of the epidermis (2C, 4C and total) and mesophyll (4C, 8C, 16C and total) tissues in the leaf. Similar to the mathematical model (See Section 4.4.2) described in Chapter 4, a simple model can be developed to predict (1) the expression of a gene in different developmental stages in the leaf and (2) the total expression in leaf tissues (Epidermis and mesophyll), as a function of endoploidy specific expression. Figures 5.5 and 5.6 illustrate some preliminary results obtained using this approach and demonstrate the potential of our predictive model for obtaining endoploidy distributions during different developmental stages of the *Arabidopsis* leaf as well. Further, the endoploidy predictor genes obtained for leaf (using a approach similar to the root model) can be used to probe the change in endoploidy distributions in leaf under various environmental stress conditions.

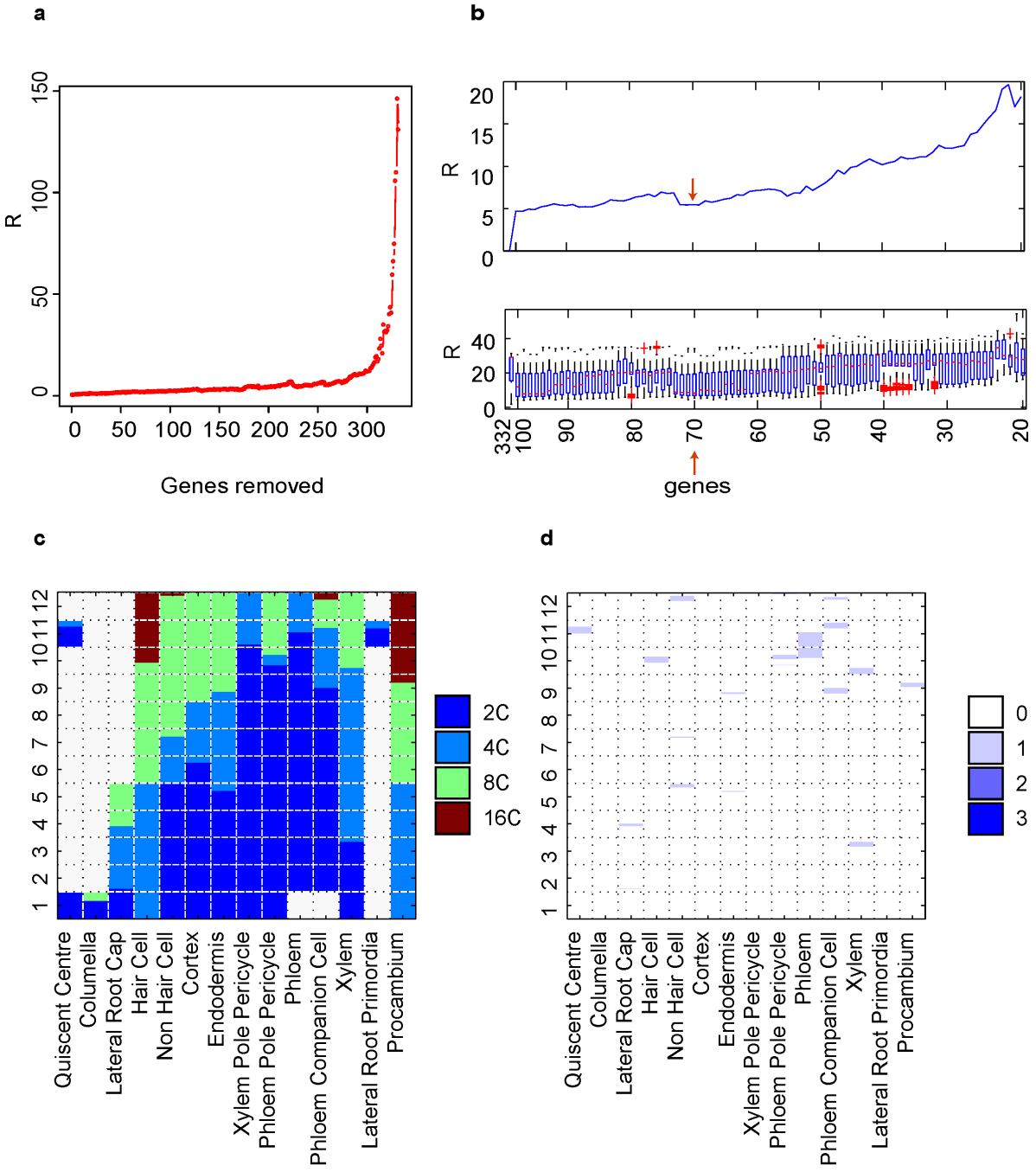


Figure 5.3: Optimised endoploidy map using a small representative set of predictor genes. (**a.**) The endoploidy border positions predicted for 332 genes were used as reference while removing one gene at a time to obtain smaller representative sets.(**b.**) The R-score line plot, which is a part of (**a**) for the set of 332 genes and sets from 100 to 20 genes) and quantile plots from R-scores obtained from 100 maps learned for individual gene set after introducing white noise (mean =0 and standard deviation = 5% of marker and slice data expression in ^{2}). The red arrow indicates the set of 70 genes, which yields a relatively stable map compared to smaller gene sets. (**c.**) Optimised endoploidy map obtained using a set of 70 genes. (**d.**) difference between the endoploidy levels at each cell on the optimised endoploidy map using 332 genes and the endoploidy map using 70 genes. The rows represent the cell types and columns represent the 12 developmental stages. The color indicates the absolute distance between the rounds of endoreplication of the same cell in both maps.

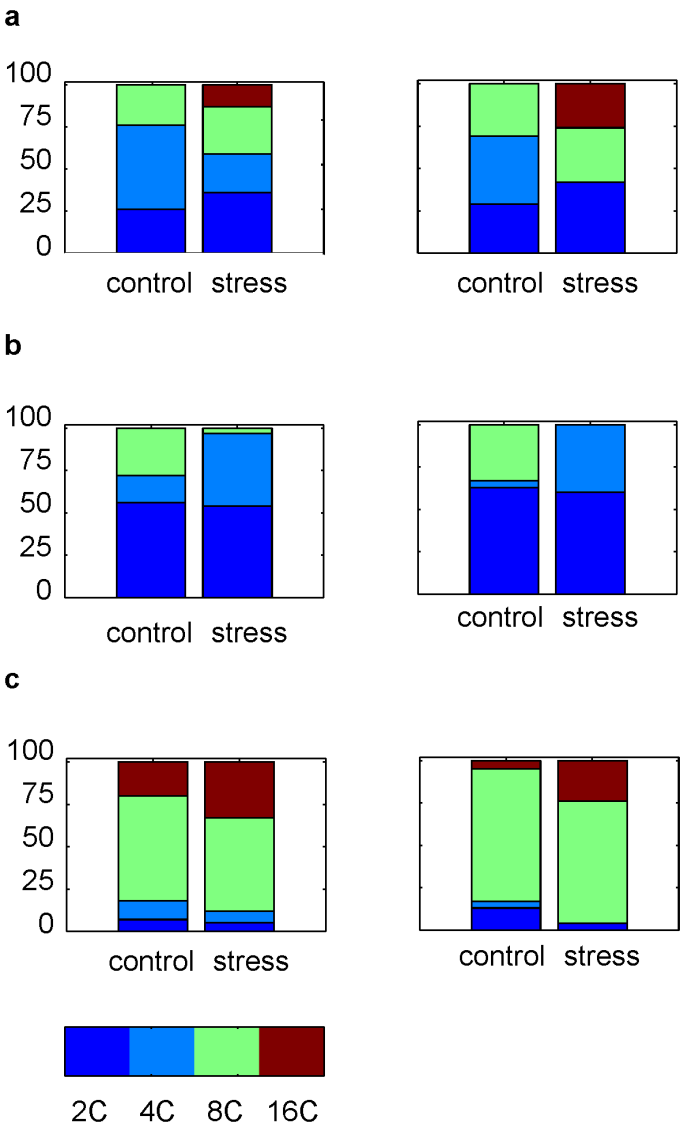


Figure 5.4: Comparison between predicted endoploidy distributions obtained using set of 323 and 70 genes. Examples of predicted endoploidy distributions under control and salt stress (140mm NaCl) for intact root (**a.**), cell type stele (WOL, see legend in Figure 4.11-c) (**b.**) and segment of root (developmental zone 4, see legend in Figure 4.12) (**c.**) using original gene set (left plot in each panel, 323 genes) and diagnostic marker gene set (right plot in each panel, 70 genes).

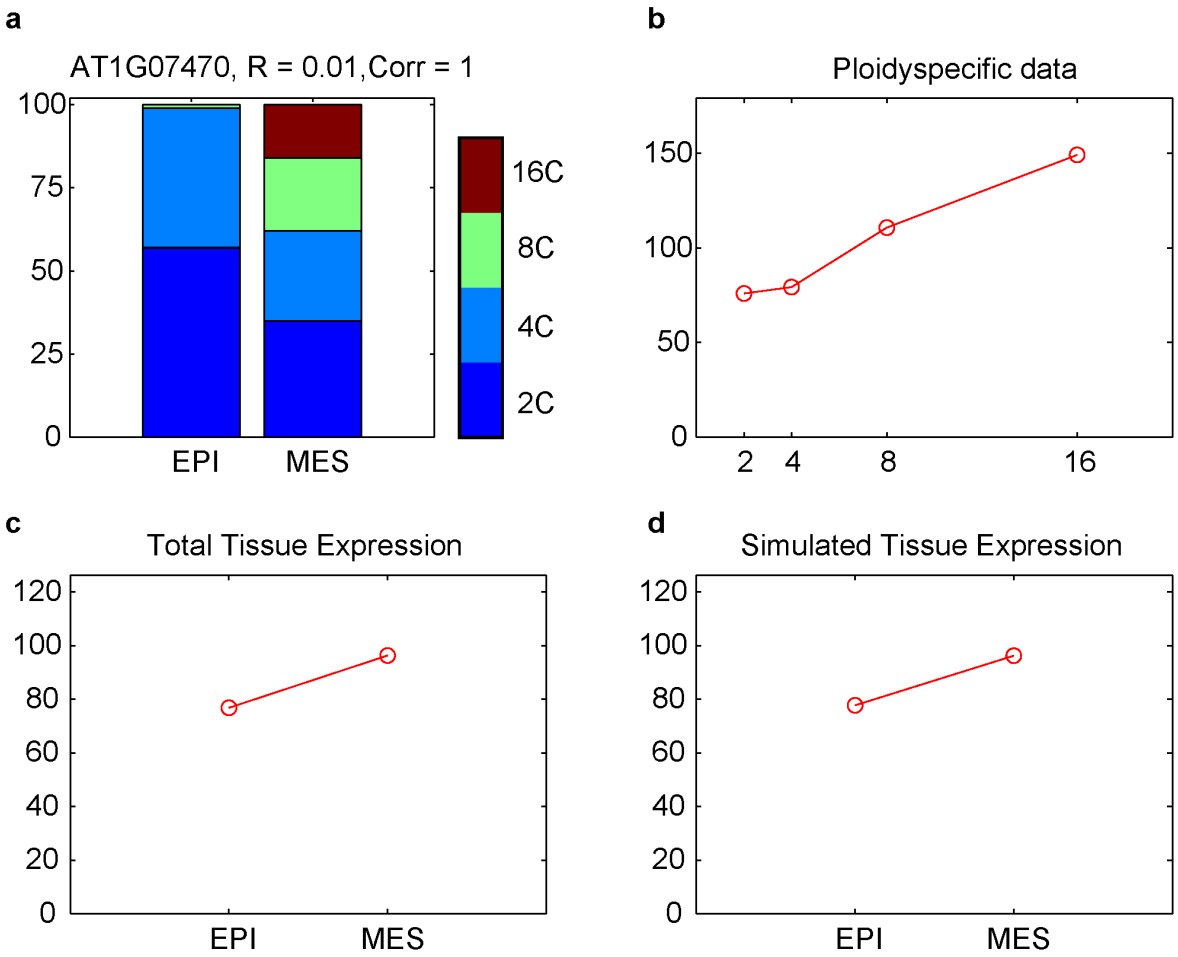


Figure 5.5: Example of the optimised expression patterns and endoploidy map learned for one gene using Model 3. (**a.**) The predicted endoploidy distributions of an *Arabidopsis* leaf tissues learned from a single gene (AT1G07470). (**b.**) The endoploidy-specific expression pattern of the gene in a. The line plot indicating measured (**c.**) and simulated (**d.**) expression patterns of the gene in epidermis and mesophyll tissues, respectively.

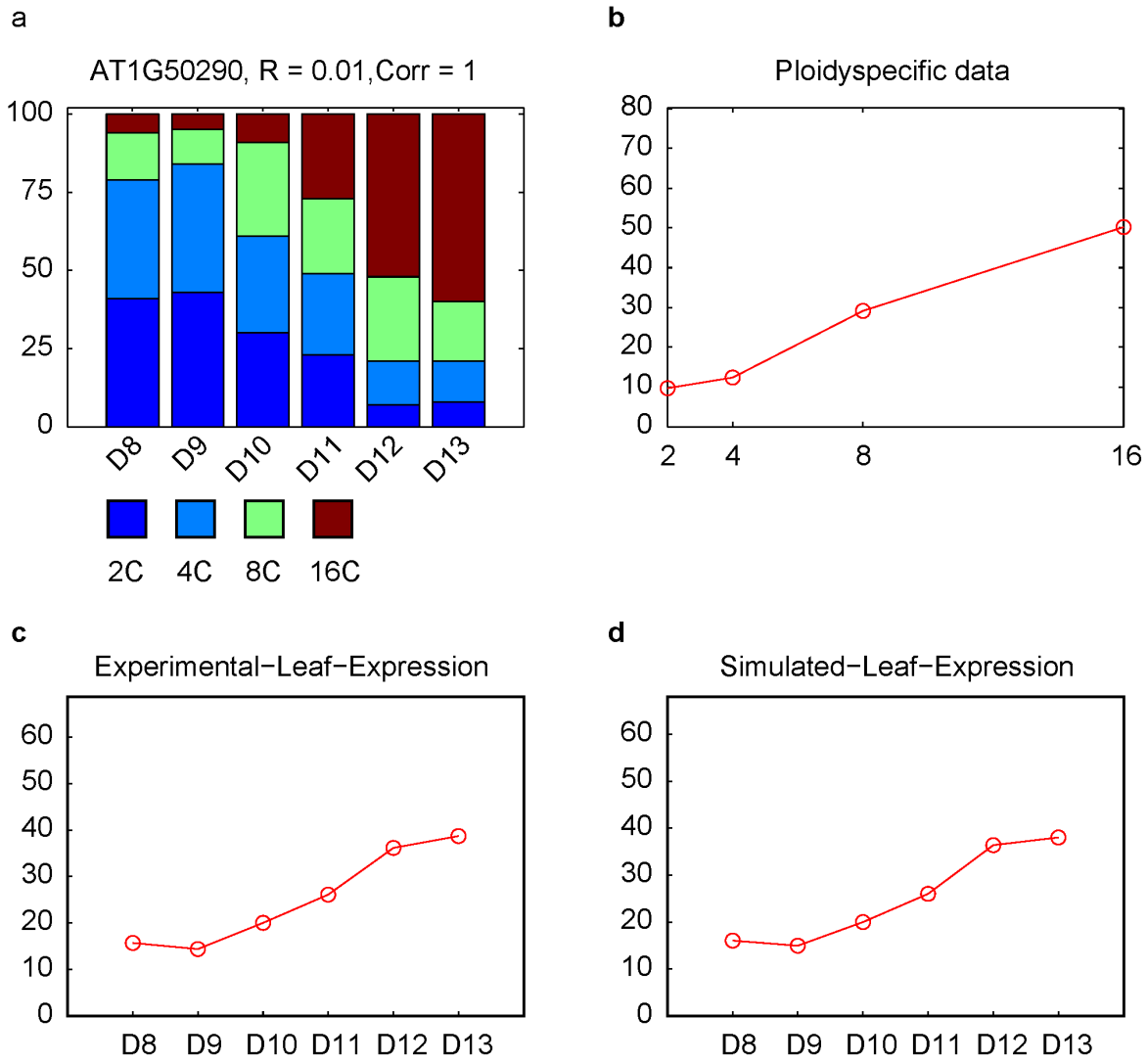


Figure 5.6: Example of the optimised expression patterns and endoploidy map learned for one gene using Model 3. (**a.**) The predicted endoploidy distributions of an *Arabidopsis* leaf tissues learned from a single gene (AT1G07470). (**b.**) The endoploidy-specific expression pattern of the gene in a. The line plot indicating measured (**c.**) and simulated (**d.**) expression patterns of the gene in epidermis and mesophyll tissues, respectively.

Potential work

As described in Chapter 4, a balanced gene set can be obtained to reliably predict endoploidy distributions in intact leaf or individual tissues under developmental and environmental cues. These obtained markers can be further investigated for their relevance in endoreplication processes using mutant line analysis. Additionally, the balanced gene sets obtained using root and leaf models can be compared to identify the organ specific as well non-specific (or global) markers involved in the endoreplication process. The current model described above uses only few data points for the predictions and does not reliably converge to the same optimisation potential over different model runs. Thus, additional data points either obtained from publicly available data or generated data need to be included for robust endoploidy predictions.

Broader significance

Leaves provide the foundation for growth of the whole plant through their photosynthetic abilities. Thus, to improve crops for yield and tolerance to environmental conditions, it is necessary to understand the effect of regulatory and environmental factors on leaf growth and development. As plant leaves employ endoreplication during their developmental program as well as in response to stress, a better understanding of this process is crucial. The potential work described here might increase our knowledge about environmental cues on endoploidy distributions in plant leaves and might identify novel components involved in this process.

5.3 Author contributions

The content of this chapter was written by myself. It resulted from the many fruitful discussions with both my promoters.

Appendices

Appendix A

Supplemental Datasets

A.1 Predicting gene function from uncontrolled expression variation among individual wild-type *Arabidopsis* plants

1. Module figures for all modules learned from the residuals dataset.
2. Excel file containing predicted GO annotations for all modules in Supplemental Dataset 1.
3. Module figures for all modules learned from the original dataset, without removing lab and accession effects.
4. Excel file containing predicted GO annotations for all modules in Supplemental Dataset 3.
5. Excel file with gene counts in the residuals and sample networks for all GO processes with at least 10 nodes in the residuals network.
6. Category-specific function prediction performance figures for all GO categories scored in Figure 2.6.

A.2 A spatio-temporal DNA endoploidy map of the *Arabidopsis* root reveals a role of the endocycle in stress adaptation

1. Excel file detailing the genes that are differentially expressed between at least two endoploidy levels.
2. Excel file containing the k-mean cluster ID of differentially expressed genes.
3. Excel file containing functional enrichment results for all endoploidy-specific gene expression clusters.
4. Expression enrichment figures for all endoploidy-specific clusters in 17 marker lines covering 14 distinct tissue types of the *Arabidopsis* root[2].
5. Expression enrichment figures for all endoploidy-specific clusters in 12 developmental stages of the *Arabidopsis* root[2].
6. Annotation information for the balanced set of 332 genes.
7. Excel file containing the GEO accession numbers and descriptions of stress and hormone treatments profiled in intact roots, root cell types and root segments as used for the Model II endoploidy distribution predictions.
8. Excel file containing reduced chi-squared statistic (R) and the Pearson correlation between simulated (using Model I) and measured expression profiles of 4378 genes.
9. Excel file containing the GEO accession numbers and descriptions of the selected stress gene expression responses profiled in intact roots, as used for pruning stress responsive genes from the set of 332 genes.
10. Endoploidy distribution prediction figures for stress/hormone treatments profiled in intact roots.

11. Endoploidy distribution prediction figures for stress/hormone treatments profiled in specific root cell types.
12. Endoploidy distribution prediction figures for stress/hormone treatments profiled in root segments.
13. Excel file containing cell numbers and their average from the QC (over 30 replicates) measured for cell type atrichoblast along the longitudinal axis of the root.

Appendix B

Academic CV

Personal information

Name	Rahul Bhosale
Address	Gordunakaai 36, 9000 Gent, Belgium
E-mail	rahul.bibb1987@gmail.com, rabho@psb.ugent.be
Phone	+32-470822538
Date of birth	10/06/1987
Place of birth	Bhatnimgaon, India

Education

2010-2014 **Doctor of Science, Biochemistry and Biotechnology**
Ghent University (UGent) / Flanders Institute for Biotechnology (VIB) - Ghent, Belgium
Funded by the Flanders Scientific Research (FWO) Grant G002911N
Supported by VIB research extension grant for PhD (04/2014-10/2014), New Phytologist Trust Travel grant (05/2013) and FEBS Youth Travel Fund (09/2010)

2004-2009 **Master of Science, Biotechnology (Five years integrated)**
University of Pune, Pune, India
Graduated with A grade
Received Summer research fellowship of the Indian Academy of Sciences

Publications

*contributed equally, #shared corresponding authors

8. **Bhosale R***, Boudolf V*, Isterdale G, Lambert G, Cuevas F, Beeckman T, Larkins J, Galbraith D, Maere S#, De Veylder L#. A spatio-temporal DNA endoploidy map of the *Arabidopsis* root reveals a role of the endocycle in stress adaptation. *In preparation for Nature*.

7. Majhi B, **Bhosale R**, Jaiwkar S, Veluthambi K (2014). Evaluation of codA, tms2, and ABRIN-A as negative selectable markers in transgenic tobacco and rice. *In Vitro Cellular & Developmental Biology - Plant*. (Accepted, DOI:10.1007/s11627-014-9625-1).

6. **Bhosale R***, **Jewell J***, Hollunder J, Koo A, Vuylsteke M, Michoel T, Hilson P, Goossens A, Howe G, Browse J, Maere S (2013). Predicting gene function from uncontrolled expression variation among individual wild-type *Arabidopsis* plants. *The Plant Cell* **25**(8):2865-2877.

5. **Bhosale R**, Rout J, Chaugule B (2012). The ethnobotanical study of an edible freshwater red alga, *Lemanea fuviatilis* (L.) C. Ag. from Manipur, India. *Ethnobotany research and applications* **10**:69-76.

4. Rout J, Chaugule B, **Bhosale R** (2011). Potential of a red alga, *Lemanea* for use as a source of biofuel. *Algae Biofuel* 111-115 (Book chapter).

3. **Bhosale R**, Chaugule B (2010). Freshwater algae as potential source of polyunsaturated fatty acids: Review. *International Journal on Algae* **20**(4): 335-356.

2. **Bhosale R**, Rajabhoj M, Chaugule B (2010). *Dunaliella* Teod as a prominent source of Eicosapentaenoic acid. *International Journal on Algae* **20**(2),193-210.

1. **Bhosale R**, Velankar D, Chaugule B (2009). Fatty acid composition of the cold-water-inhabiting freshwater red alga *Sirodotia* Kylin. *Journal of Applied Phycology* **21**: 99–102.

Selected meetings

Interdisciplinary Plant group 30th Annual symposium on Root Biology 2013, (IPG2013) - Missouri, USA
Poster: A virtual endoploidy map of *Arabidopsis* root.

VIB Annual Seminar 2013 (VIB2013) - Blankenberge, Belgium
Oral presentation: A virtual endoploidy map of *Arabidopsis* root.

International Conference on Systems Biology 2012 (ICSB2012) - Toronto, Canada
Poster: Analysis of gene residual expression data from individual *Arabidopsis* plants: potential and implications.

Frontiers in plant biology: From discovery to applications 2012 - Ghent, Belgium
Poster: A virtual endoploidy map of *Arabidopsis* root.

The Plant Growth Biology and Modeling symposium 2012 (PGBM2011) - Elche, Spain.
Oral presentation: Residual uncontrolled variation in a tightly controlled gene expression experiment on individual *Arabidopsis* plants yields biologically relevant expression modules.

FEBS Advanced Course in Analysis and Engineering of Biomolecular Systems 2010 - Spetses, Greece
Poster: Residual uncontrolled variation in a tightly controlled gene expression experiment on individual *Arabidopsis* plants yields biologically relevant expression modules.

Selected training - Transferable skills courses

Training program on effective oral presentations 2013 - Ghent, Belgium
VIB research training course

Plant circuit dynamics 2013 - Ghent, Belgium
VIB research training course

Effective writing for Life Sciences Research 2013 - Ghent, Belgium
VIB research training course

Advanced Academic English: Writing Skills 2013 - Ghent, Belgium
Ghent University doctoral schools course

Advanced Academic English: Conference Skills 2013- Ghent, Belgium

Ghent University doctoral schools course

Appendix C

Bibliography

[1] Maere S, Van Dijck P, and Kuiper M 2008. Extracting expression modules from perturbational gene expression compendia. *BMC Systems Biology* **2**(1):33.

[2] Brady SM, Orlando DA, Lee JY, Wang JY, Koch J, et al. 2007. A High-Resolution Root Spatiotemporal Map Reveals Dominant Expression Patterns. *Science* **318**(5851):801–806.

[3] Schilling M, Pfeifer AC, Bohl S, and Klingmuller U 2008. Standardizing experimental protocols. *Current Opinion in Biotechnology* **19**(4):354–359.

[4] Massonnet C, Vile D, Fabre J, Hannah MA, Caldana C, et al. 2010. Probing the reproducibility of leaf growth and molecular phenotypes: a comparison of three Arabidopsis accessions cultivated in ten laboratories. *Plant physiology* **152**(4):2142–2157.

[5] Watson JD and Crick FHC 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**(4356):737–738.

[6] Crick FH 1958. On protein synthesis. *Symposia of the Society for Experimental Biology* **12**:138.

[7] Crick F 1970. Central dogma of molecular biology. *Nature* **227**(5258):561–563.

[8] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, et al. 2002. Protein Function. In *Molecular Biology of the Cell*. Garland Science, New York.

[9] Initiative TAG 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**(6814):796–815.

[10] Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, et al. 2005. A gene expression map of Arabidopsis thaliana development. *Nature Genetics* **37**(5):501–506.

[11] Eubel H, Meyer EH, Taylor NL, Bussell JD, O'Toole N, et al. 2008. Novel proteins, putative membrane transporters, and an integrated metabolic network are revealed by quantitative proteomic analysis of Arabidopsis cell culture peroxisomes. *Plant physiology* **148**(4):1809–1829.

[12] Atkinson MR, Savageau MA, Myers JT, and Ninfa AJ 2003. Development of Genetic Circuitry Exhibiting Toggle Switch or Oscillatory Behavior in Escherichia coli. *Cell* **113**(5):597–607.

[13] Covert MW, Knight EM, Reed JL, Herrgard MJ, and Palsson BO 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature Cell Biology* **429**(6987):92–96.

[14] Ozbudak EM, Thattai M, Lim HN, Shraiman BI, and van Oudenaarden A 2004. Multistability in the lactose utilization network of Escherichia coli. *Nature* **427**(6976):737–740.

[15] Park H, Pontius W, Guet CC, Marko JF, Emonet T, et al. 2010. Interdependence of behavioural variability and response to small stimuli in bacteria. *Nature* **468**(7325):819–823.

[16] Rosenfeld N, Young JW, Alon U, Swain PS, and Elowitz MB 2005. Gene Regulation at the Single-Cell Level. *Science* **307**(5717):1962–1965.

[17] Alwine JC, Kemp DJ, and Stark GR 1977. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America* **74**(12):5350–5354.

[18] Schena M, Shalon D, Davis RW, and Brown PO 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235):467–470.

[19] Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**(13):1675–1680.

[20] Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, et al. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques* **45**(1):81–94.

[21] Bustin SA 2000. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of molecular endocrinology* **25**(2):169–193.

[22] Velculescu VE, Zhang L, Vogelstein B, and Kinzler KW 1995. Serial analysis of gene expression. *Science* **270**(5235):484–487.

[23] Towbin H, Staehelin T, and Gordon J 1979. Electrophoretic Transfer of Proteins From Polyacrylamide Gels to Nitrocellulose Sheets - Procedure and Some Applications. *Proceedings of the National Academy of Sciences of the United States of America* **76**(9):4350–4354.

[24] Weston AD and Hood L 2004. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *Journal of proteome research* **3**(2):179–196.

[25] Mutz KO, Heilkenbrinker A, Lönne M, Walter JG, and Stahl F 2013. Transcriptome analysis using next-generation sequencing. *Current Opinion in Biotechnology* **24**(1):22–30.

[26] Zhao S, Fung-Leung WP, Bittner A, Ngo K, and Liu X 2014. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS one* **9**(1):e78644.

[27] Redman JC, Haas BJ, Tanimoto G, and Town CD 2004. Development and evaluation of an Arabidopsiswhole genome Affymetrix probe array. *The Plant Journal* **38**(3):545–561.

[28] Rehrauer H, Aquino C, Gruissem W, Henz SR, Hilson P, et al. 2010. AGRONOMICS1: A New Resource for Arabidopsis Transcriptome Profiling. *Plant physiology* **152**(2):487–499.

[29] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)* **4**(2):249–264.

[30] Bolstad BM, Irizarry RA, Astrand M, and Speed TP 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**(2):185–193.

[31] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**(4):e15.

[32] Bland JM and Altman DG 1995. Multiple significance tests: the Bonferroni method. *Bmj* **310**(6973):170.

[33] Benjamini Y and Hochberg Y 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 289–300.

[34] Smyth GK 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3**:Article3.

[35] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**(10).

[36] Consortium GO 2006. The gene ontology (GO) project in 2006. *Nucleic Acids Research* **34**(suppl 1):D322–D326.

[37] Huang DW, Sherman BT, and Lempicki RA 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**(1):1–13.

[38] Beißbarth T and Speed TP 2004. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**(9):1464–1465.

[39] Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, et al. 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology* **4**(4):R28.

[40] Maere S, Heymans K, and Kuiper M 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**(16):3448–3449.

[41] Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, et al. 2007. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology* **8**(9):R183.

[42] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**(43):15545–15550.

[43] Kim SY and Volsky DJ 2005. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* **6**:144.

[44] Smid M and Dorssers LCJ 2004. GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. *Bioinformatics* **20**(16):2618–2625.

[45] Nam D, Kim SB, Kim SK, Yang S, Kim SY, et al. 2006. ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics* **22**(18):2249–2253.

[46] Bauer S, Grossmann S, Vingron M, and Robinson PN 2008. Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24**(14):1650–1651.

[47] Martin D, Brun C, Remy E, Mouren P, Thieffry D, et al. 2004. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology* **5**(12):R101.

[48] Walker MG 1999. Prediction of Gene Function by Genome-Scale Expression Analysis: Prostate Cancer-Associated Genes. *Genome Research* **9**(12):1198–1203.

[49] Ng SK, Zhu Z, and Ong YS 2004. Whole-genome functional classification of genes by latent semantic analysis on microarray data. *Proceedings of the second conference on Asia-Pacific bioinformatics* **29**:123–129.

[50] Kim DW, Lee KH, and Lee D 2005. Detecting clusters of different geometrical shapes in microarray gene expression data. *Bioinformatics* **21**(9):1927–1934.

[51] Eisen MB, Spellman PT, Brown PO, and Botstein D 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**(25):14863–14868.

[52] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, and Church GM 1999. Systematic determination of genetic network architecture. *Nature Genetics* **22**(3):281–285.

[53] Lukashin AV and Fuchs R 2001. Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics* **17**(5):405–414.

[54] Xu Y, Olman V, and Xu D 2002. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics* **18**(4):536–545.

[55] Califano A, Stolovitzky G, and Tu Y 2000. Analysis of gene expression microarrays for phenotype classification. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **8**:75–85.

[56] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, et al. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America* **97**(1):262–267.

[57] Mateos A, Dopazo J, Jansen R, Tu Y, Gerstein M, et al. 2002. Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Research* **12**(11):1703–1715.

[58] Li XL, Tan YC, and Ng SK 2006. Systematic gene function prediction from gene expression data by using a fuzzy nearest-cluster method. *BMC Bioinformatics* **7**(Suppl 4):S23.

[59] Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, et al. 2006. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **22**(9):1122–1129.

[60] Madeira SC and Oliveira AL 2004. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **1**(1):24–45.

[61] Oghabian A, Kilpinen S, Hautaniemi S, and Czeizler E 2014. Biclustering methods: biological relevance and application in gene expression analysis. *PloS one* **9**(3):e90801.

[62] Teng L and Chan L 2008. Discovering Biclusters by Iteratively Sorting with Weighted Correlation Coefficient in Gene Expression Data. *Journal of Signal Processing Systems* **50**(3):267–280–280.

[63] Ayadi W, Elloumi M, and Hao JK 2009. A biclustering algorithm based on a bicluster enumeration tree: application to DNA microarray data. *BioData mining* **2**:9.

[64] Cheng Y and Church GM 2000. Biclustering of expression data. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **8**:93–103.

[65] Yang J, Wang H, Wang W, and Yu PS 2005. An improved biclustering method for analyzing gene expression profiles. *International Journal on Artificial Intelligence Tools* **14**(05):771–789.

[66] Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, et al. 2006. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **22**(9):1122–1129.

[67] Kluger Y, Basri R, Chang JT, and Gerstein M 2003. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research* **13**(4):703–716.

[68] Murali TM and Kasif S 2003. Extracting conserved gene expression motifs from gene expression data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* pages 77–88.

[69] Getz G, Levine E, and Domany E 2000. Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the United States of America* **97**(22):12079–12084.

[70] Ihmels J, Bergmann S, and Barkai N 2004. Defining transcription modules using large-scale gene expression data. *Bioinformatics* **20**(13):1993–2003.

[71] Tang C, Zhang L, Zhang A, and Ramanathan M 2001. Interrelated two-way clustering: an unsupervised approach for gene expression data analysis pages 41–48.

[72] Reiss DJ, Baliga NS, and Bonneau R 2006. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* **7**:280.

[73] Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, et al. 2010. FABIA: factor analysis for bicluster acquisition. *Bioinformatics* **26**(12):1520–1527.

[74] Sheng Q, Moreau Y, and De Moor B 2003. Biclustering microarray data by Gibbs sampling. *Bioinformatics* **19 Suppl 2**:ii196–205.

[75] Lazzeroni L and Owen A 2002. Plaid models for gene expression data. *Statistica sinica* **12**(1):61–86.

[76] Tanay A, Sharan R, and Shamir R 2002. Discovering statistically significant biclusters in gene expression data. *Bioinformatics* **18 Suppl 1**:S136–44.

[77] Sheng Q, Moreau Y, and De Moor B 2003. Biclustering microarray data by Gibbs sampling. *Bioinformatics* **19 Suppl 2**:ii196–205.

[78] Stuart JM, Segal E, Koller D, and Kim SK 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* **302**(5643):249–255.

[79] Hegyi H and Gerstein M 1999. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *Journal of molecular biology* **288**(1):147–164.

[80] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**(6770):623–627.

[81] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, and Yeates TO 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences* **96**(8):4285–4288.

[82] Mostafavi S, Ray D, Warde-Farley D, Grouios C, and Morris Q 2008. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology* **9 Suppl 1**:S4.

[83] Tian W, Zhang LV, Taşan M, Gibbons FD, King OD, et al. 2008. Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function. *Genome Biology* **9 Suppl 1**:S7.

[84] Troyanskaya OG, Dolinski K, Owen AB, Altman RB, and Botstein D 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences of the United States of America* **100**(14):8348–8353.

[85] Sharan R, Ulitsky I, and Shamir R 2007. Network-based prediction of protein function. *Molecular Systems Biology* **3**:88.

[86] Hwang S, Rhee SY, Marcotte EM, and Lee I 2011. Systematic prediction of gene function in Arabidopsis thaliana using a probabilistic functional gene network. *Nature protocols* **6**(9):1429–1442.

[87] Pop A, Huttenhower C, Iyer-Pascuzzi A, Benfey PN, and Troyanskaya OG 2010. Integrated functional networks of process, tissue, and developmental stage specific interactions in Arabidopsis thaliana. *BMC Systems Biology* **4**(1):180.

[88] Smoot M, Ono K, Ideker T, and Maere S 2011. PiNGO: a Cytoscape plugin to find candidate genes in biological networks. *Bioinformatics* **27**(7):1030–1031.

[89] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**(11):2498–2504.

[90] Ideker T, Galitski T, and Hood L 2001. A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics* **2**:343–372.

[91] Richter SH, Garner JP, Zipser B, Lewejohann L, Sachser N, et al. 2011. Effect of Population Heterogenization on the Reproducibility of Mouse Behavior: A Multi-Laboratory Study. *PloS one* **6**(1).

[92] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**(1):25–29.

[93] Deng MH, Tu ZD, Sun FZ, and Chen T 2004. Mapping gene ontology to proteins based on protein-protein interaction data. *Bioinformatics* **20**(6):895–902.

[94] Lamesch P, Berardini TZ, Li DH, Swarbreck D, Wilks C, et al. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* **40**(D1):D1202–D1210.

[95] Mugford SG, Yoshimoto N, Reichelt M, Wirtz M, Hill L, et al. 2009. Disruption of Adenosine-5 '-Phosphosulfate Kinase in Arabidopsis Reduces Levels of Sulfated Secondary Metabolites. *The Plant Cell* **21**(3):910–927.

[96] Bu Q, Jiang H, Li CB, Zhai Q, Zhang J, et al. 2008. Role of the Arabidopsis thaliana NAC transcription factors ANAC019 and ANAC055 in regulating jasmonic acid-signaled defense responses. *Cell Research* **18**(7):756–767.

[97] Reinbothe C, Springer A, Samol I, and Reinbothe S 2009. Plant oxylipins: role of jasmonic acid during programmed cell death, defence and leaf senescence. *FEBS Journal* **276**(17):4666–4681.

[98] Heyndrickx KS and Vandepoele K 2012. Systematic Identification of Functional Plant Modules through the Integration of Complementary Data Sources. *Plant physiology* **159**(3):884–901.

[99] Bartel B and Fink GR 1995. Ilr1, an Amidohydrolase That Releases Active Indole-3-Acetic-Acid from Conjugates. *Science* **268**(5218):1745–1748.

[100] Davies RT, Goetz DH, Lasswell J, Anderson MN, and Bartel B 1999. IAR3 encodes an auxin conjugate hydrolase from Arabidopsis. *The Plant Cell* **11**(3):365–376.

[101] Woldemariam MG, Onkokesung N, Baldwin IT, and Galis I 2012. Jasmonoyl-l-isoleucine hydrolase 1 (JIH1) regulates jasmonoyl-l-isoleucine levels and attenuates plant defenses against herbivores. *The Plant journal : for cell and molecular biology* **72**(5):758–767.

[102] Kodaira KS, Qin F, Tran LSP, Maruyama K, Kidokoro S, et al. 2011. Arabidopsis Cys2/His2 Zinc-Finger Proteins AZF1 and AZF2 Negatively Regulate Abscisic Acid-Repressive and Auxin-Inducible Genes under Abiotic Stress Conditions. *Plant physiology* **157**(2):742–756.

[103] Zander M, La Camera S, Lamotte O, Métraux JP, and Gatz C 2009. Arabidopsis thaliana class-II TGA transcription factors are essential activators of jasmonic acid/ethylene-induced defense responses. *The Plant Journal* **61**(2):200–210.

[104] Yoo SD, Cho YH, Tena G, Xiong Y, and Sheen J 2008. Dual control of nuclear EIN3 by bifurcate MAPK cascades in C2H4 signalling. *Nature* **451**(7180):789–795.

[105] Zhu Z, An F, Feng Y, Li P, Xue L, et al. 2011. Derepression of ethylene-stabilized transcription factors (EIN3/EIL1) mediates jasmonate and ethylene signaling synergy in Arabidopsis. *Proceedings of the National Academy of Sciences* **108**(30):12539–12544.

[106] Jia XY, Xu CY, Jing RL, Li RZ, Mao XG, et al. 2008. Molecular cloning and characterization of wheat calreticulin (CRT) gene involved in drought-stressed responses. *Journal of experimental botany* **59**(4):739–751.

[107] Wan D, Li R, Zou B, Zhang X, Cong J, et al. 2012. Calmodulin-binding protein CBP60g is a positive regulator of both disease resistance and drought tolerance in Arabidopsis. *Plant Cell Reports* **31**(7):1269–1281.

[108] Chen H, Lai Z, Shi J, Xiao Y, Chen Z, et al. 2010. Roles of arabidopsis WRKY18, WRKY40 andWRKY60 transcription factors in plant responsesto abscisic acid and abiotic stress. *BMC Plant Biology* **10**(1):281.

[109] Persak H and Pitzschke A 2013. Tight Interconnection and Multi-Level Control of Arabidopsis MYB44 in MAPK Cascade Signalling. *PloS one* **8**(2):e57547.

[110] Xu J, Li Y, Wang Y, Liu H, Lei L, et al. 2008. Activation of MAPK Kinase 9 Induces Ethylene and Camalexin Biosynthesis and Enhances Sensitivity to Salt Stress in Arabidopsis. *Journal of Biological Chemistry* **283**(40):26996–27006.

[111] Birkenbihl RP, Diezel C, and Somssich IE 2012. Arabidopsis WRKY33 Is a Key Transcriptional Regulator of Hormonal and Metabolic Responses toward Botrytis cinerea Infection. *Plant physiology* **159**(1):266–285.

[112] Berrocal-Lobo M, Molina A, and Solano R 2002. Constitutive expression of ETHYLENE-RESPONSE-FACTOR1 in Arabidopsis confers resistance to several necrotrophic fungi. *The Plant journal : for cell and molecular biology* **29**(1):23–32.

[113] Thomma BP, Nelissen I, Eggermont K, and Broekaert WF 1999. Deficiency in phytoalexin production causes enhanced susceptibility of Arabidopsis thaliana to the fungus Alternaria brassicicola. *The Plant journal : for cell and molecular biology* **19**(2):163–171.

[114] Henriksson E and Nordin Henriksson K 2005. Salt-stress signalling and the role of calcium in the regulation of the Arabidopsis ATHB7 gene. *Plant, Cell & Environment* **28**:202–210.

[115] Baldwin IT 1998. Jasmonate-induced responses are costly but benefit plants under attack in native populations. *Proceedings of the National Academy of Sciences* **95**(14):8113–8118.

[116] Dunlop MJ, Cox RS, Levine JH, Murray RM, and Elowitz MB 2008. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature Genetics* **40**(12):1493–1498.

[117] Munsky B, Neuert G, and van Oudenaarden A 2012. Using Gene Expression Noise to Understand Gene Regulation. *Science* **336**(6078):183–187.

[118] Stewart-Ornstein J, Weissman JS, and El-Samad H 2012. Cellular Noise Regulons Underlie Fluctuations in Saccharomyces cerevisiae. *Molecular Cell* **45**(4):483–493.

[119] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102**(1):109–126.

[120] Chua G, Robinson MD, Morris Q, and Hughes TR 2004. Transcriptional networks: reverse-engineering gene regulation on a global scale. *Current Opinion in Microbiology* **7**(6):638–646.

[121] Ma S and Bohnert HJ 2007. Integration of Arabidopsis thaliana stress-related transcript profiles, promoter structures, and cell-specific expression. *Genome Biology* **8**(4):R49.

[122] He F, Balling R, and Zeng AP 2009. Reverse engineering and verification of gene networks: Principles, assumptions, and limitations of present methods and future perspectives. *Journal of Biotechnology* **144**(3):190–203.

[123] Lee I, Ambaru B, Thakkar P, Marcotte EM, and Rhee SY 2010. Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nature Biotechnology* **28**(2):149–U14.

[124] Kliebenstein D 2009. Quantitative Genomics: Analyzing Intraspecific Variation Using Global Gene Expression Polymorphisms or eQTLs. *Annual Review of Plant Biology* **60**:93–114.

[125] Nayak RR, Kearns M, Spielman RS, and Cheung VG 2009. Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome Research* **19**(11):1953–1962.

[126] Chan EKF, Rowe HC, Corwin JA, Joseph B, and Kliebenstein DJ 2011. Combining Genome-Wide Association Mapping and Transcriptional Networks to Identify Novel Genes Controlling Glucosinolates in Arabidopsis thaliana. *PLoS Biology* **9**(8).

[127] Cubillos FA, Coustham V, and Loudet O 2012. Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. *Current Opinion in Plant Biology* **15**(2):192–198.

[128] Weigel D 2012. Natural Variation in Arabidopsis: From Molecular Genetics to Ecological Genomics. *Plant physiology* **158**(1):2–22.

[129] Nagano AJ, Sato Y, Mihara M, Antonio BA, Motoyama R, et al. 2012. Deciphering and Prediction of Transcriptome Dynamics under Fluctuating Field Conditions. *Cell* **151**(6):1358–1369.

[130] Richards CL, Rosas U, Banta J, Bhambhra N, and Purugganan MD 2012. Genome-Wide Patterns of Arabidopsis Gene Expression in Nature. *PLoS genetics* **8**(4):482–495.

[131] De Bodt S, Carvajal D, Hollunder J, Van den Cruyce J, Movahedi S, et al. 2010. CORNET: A User-Friendly Tool for Data Mining and Integration. *Plant physiology* **152**(3):1167–1179.

[132] Casneuf T, Van de Peer Y, and Huber W 2007. In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics* **8**(1):461.

[133] Smoot M, Ono K, Ideker T, and Maere S 2011. PiNGO: a Cytoscape plugin to find candidate genes in biological networks. *Bioinformatics* **27**(7):1030–1031.

[134] Koo AJK, Gao XL, Jones AD, and Howe GA 2009. A rapid wound signal activates the systemic synthesis of bioactive jasmonates in Arabidopsis. *The Plant journal : for cell and molecular biology* **59**(6):974–986.

[135] Koo AJK, Cooke TF, and Howe GA 2011. Cytochrome P450 CYP94B3 mediates catabolism and inactivation of the plant hormone jasmonoyl-L-isoleucine. *Proceedings of the National Academy of Sciences of the United States of America* **108**(22):9298–9303.

[136] Farmer EE, Johnson RR, and Ryan CA 1992. Regulation of Expression of Proteinase-Inhibitor Genes by Methyl Jasmonate and Jasmonic Acid. *Plant physiology* **98**(3):995–1002.

[137] Fonseca S, Chini A, Hamberg M, Adie B, Porzel A, et al. 2009. (+)-7-iso-Jasmonoyl-L-isoleucine is the endogenous bioactive jasmonate. *Nature Chemical Biology* **5**(5):344–350.

[138] Suza WP, Rowe ML, Hamberg M, and Staswick PE 2010. A tomato enzyme synthesizes (+)-7-iso-jasmonoyl-l-isoleucine in wounded leaves. *Planta* **231**(3):717–728.

[139] Schneider CA, Rasband WS, and Eliceiri KW 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**(7):671–675.

[140] Inzé D and De Veylder L 2006. Cell cycle regulation in plant development. *Annual review of genetics* **40**:77–105.

[141] De Veylder L, Beeckman T, and Inzé D 2007. The ins and outs of the plant cell cycle. *Nature Reviews Molecular Cell Biology* **8**(8):655–665.

[142] GALBRAITH DW, HARKINS KR, and KNAPP S 1991. Systemic Endopolyploidy in Arabidopsis-Thaliana. *Plant physiology* **96**(3):985–989.

[143] Gutierrez C 2009. The Arabidopsis Cell Division Cycle. *The Arabidopsis book / American Society of Plant Biologists* **7**:e0120.

[144] Beemster GTS 2005. Genome-Wide Analysis of Gene Expression Profiles Associated with Cell Cycle Transitions in Growing Organs of Arabidopsis. *Plant physiology* **138**(2):734–743.

[145] Vandepoele K, Vlieghe K, Florquin K, Hennig L, Beemster GT, et al. 2005. Genome-wide identification of potential plant E2F target genes. *Plant physiology* **139**(1):316–328.

[146] Naouar N, Vandepoele K, Lammens T, Casneuf T, Zeller G, et al. 2009. Quantitative RNA expression analysis with Affymetrix Tiling 1.0R arrays identifies new E2F target genes. *The Plant Journal* **57**(1):184–194.

[147] Berckmans B, Lammens T, Van Den Daele H, Magyar Z, Bögre L, et al. 2011. Light-dependent regulation of DEL1 is determined by the antagonistic action of E2Fb and E2Fc. *Plant physiology* **157**(3):1440–1451.

[148] Boudolf V 2004. The Plant-Specific Cyclin-Dependent Kinase CDKB1;1 and Transcription Factor E2Fa-DPa Control the Balance of Mitotically Dividing and Endoreduplicating Cells in Arabidopsis. *The Plant Cell* **16**(10):2683–2692.

[149] del Pozo JC 2006. The Balance between Cell Division and Endoreplication Depends on E2FC-DPB, Transcription Factors Regulated by the Ubiquitin-SCFSKP2A Pathway in Arabidopsis. *The Plant Cell* **18**(9):2224–2235.

[150] Lammens T, Boudolf V, Kheibarshekan L, Panagiotis Zalmas L, Gaamouche T, et al. 2008. Atypical E2F activity restrains APC/CCCS52A2 function obligatory for endocycle onset. *Proceedings of the National Academy of Sciences* **105**(38):14721–14726.

[151] Ito M, Iwase M, Kodama H, Lavisse P, Komamine A, et al. 1998. A novel cis-acting element in promoters of plant B-type cyclin genes activates M phase-specific transcription. *Plant Cell* **10**(3):331–341.

[152] Lin Z, Yin K, Zhu D, Chen Z, Gu H, et al. 2007. AtCDC5 regulates the G2 to M transition of the cell cycle and is critical for the function of Arabidopsis shoot apical meristem. *Cell Research* **17**(9):815–828.

[153] Petroni K, Falasca G, Calvenzani V, Allegra D, Stolfi C, et al. 2008. The AtMYB11 gene from Arabidopsis is expressed in meristematic cells and modulates growth in planta and organogenesis in vitro. *Journal of experimental botany* **59**(6):1201–1213.

[154] Ito M, Araki S, Matsunaga S, Itoh T, Nishihama R, et al. 2001. G2/M-Phase-Specific transcription during the plant cell cycle is mediated by c-Myb-like transcription factors (vol 13, pg 1891, 2001). *Plant Cell* **13**(9):2159–2159.

[155] Kato K, Galis I, Suzuki S, Araki S, Demura T, et al. 2009. Preferential Up-Regulation of G2/M Phase-Specific Genes by Overexpression of the Hyperactive Form of NtmybA2 Lacking Its Negative Regulation Domain in Tobacco BY-2 Cells. *Plant physiology* **149**(4):1945–1957.

[156] Kasili R, Walker JD, Simmons LA, Zhou J, De Veylder L, et al. 2010. SIAMESE cooperates with the CDH1-like protein CCS52A1 to establish endoreplication in Arabidopsis thaliana trichomes. *Genetics* **185**(1):257–268.

[157] Boudolf V, Lammens T, Boruc J, Van Leene J, Van den Daele H, et al. 2009. CDKB1;1 Forms a Functional Complex with CYCA2;3 to Suppress Endocycle Onset. *Plant physiology* **150**(3):1482–1493.

[158] Mathieu-Rivet E, Gévaudant F, Sicard A, Salar S, Do PT, et al. 2010. Functional analysis of the anaphase promoting complex activator CCS52A highlights the crucial role of endo-reduplication for fruit growth in tomato. *The Plant journal : for cell and molecular biology* **62**(5):727–741.

[159] Cebolla A, Vinardell JM, Kiss E, Oláh B, Roudier F, et al. 1999. The mitotic inhibitor ccs52 is required for endoreduplication and ploidy-dependent cell enlargement in plants. *The EMBO journal* **18**(16):4476–4484.

[160] Walker JD, Oppenheimer DG, Concienne J, and Larkin JC 2000. SIAMESE, a gene controlling the endoreduplication cell cycle in Arabidopsis thaliana trichomes. *Development* **127**(18):3931–3940.

[161] Churchman ML, Brown ML, Kato N, Kirik V, Hulskamp M, et al. 2006. SIAMESE, a plant-specific cell cycle regulator, controls endoreplication onset in Arabidopsis thaliana. *Plant Cell* **18**(11):3145–3157.

[162] Schnittger A, Schöbinger U, Bouyer D, Weinl C, Stierhof YD, et al. 2002. Ectopic D-type cyclin expression induces not only DNA replication but also cell division in Arabidopsis trichomes. *Proceedings of the National Academy of Sciences of the United States of America* **99**(9):6410–6415.

[163] Dewitte W, Scofield S, Alcasabas AA, Maughan SC, Menges M, et al. 2007. Arabidopsis CYCD3 D-type cyclins link cell proliferation and endocycles and are rate-limiting for cytokinin responses. *Proceedings of the National Academy of Sciences of the United States of America* **104**(36):14537–14542.

[164] Dissmeyer N, Nowack MK, Pusch S, Stals H, Inzé D, et al. 2007. T-loop phosphorylation of Arabidopsis CDKA;1 is required for its function and can be partially substituted by an aspartate residue. *Plant Cell* **19**(3):972–985.

[165] Menges M, De Jager SM, Gruissem W, and Murray JAH 2005. Global analysis of the core cell cycle regulators of Arabidopsis identifies novel genes, reveals multiple and highly specific profiles of expression and provides a coherent model for plant cell cycle control. *The Plant journal : for cell and molecular biology* **41**(4):546–566.

[166] Verkest A 2005. The Cyclin-Dependent Kinase Inhibitor KRP2 Controls the Onset of the Endoreduplication Cycle during Arabidopsis Leaf Development through Inhibition of Mitotic CDKA;1 Kinase Complexes. *The Plant Cell* **17**(6):1723–1736.

[167] Wang H, Zhou Y, and Fowke LC 2006. The emerging importance of cyclin-dependent kinase inhibitors in the regulation of the plant cell cycle and related processesThis review is one of a selection of papers published in the Special Issue on Plant Cell Biology. *Canadian Journal of Botany* **84**(4):640–650.

[168] Flemming AJ, Shen ZZ, Cunha A, Emmons SW, and Leroi AM 2000. Somatic polyploidization and cellular proliferation drive body size evolution in nematodes. *Proceedings of the National Academy of Sciences of the United States of America* **97**(10):5285–5290.

[169] Ganot P and Thompson EM 2002. Patterning through Differential Endoreduplication in Epithelial Organogenesis of the Chordate, Oikopleura dioica. *Developmental biology* **252**(1):59–71.

[170] Audibert A, Simon F, and Gho M 2005. Cell cycle diversity involves differential regulation of Cyclin E activity in the Drosophila bristle cell lineage. *Development* **132**(10):2287–2297.

[171] Unhavaithaya Y and Orr-Weaver TL 2012. Polyploidization of glia in neural development links tissue growth to blood-brain barrier integrity. *Genes & Development* **26**(1):31–36.

[172] Hammond MP and Laird CD 1985. Chromosome structure and DNA replication in nurse and follicle cells of Drosophila melanogaster. *Chromosoma* **91**(3-4):267–278.

[173] Hammond MP and Laird CD. Control of DNA replication and spatial distribution of defined DNA sequences in salivary gland cells of Drosophila melanogaster. *Chromosoma* **91**(3-4):279–286.

[174] Gregory TR and Shorthouse DP 2003. Genome sizes of spiders. *The Journal of heredity* **94**(4):285–290.

[175] Mandrioli M, Mola L, Cuoghi B, and Sonetti D 2010. Endoreplication: a molecular trick during animal neuron evolution. *The Quarterly review of biology* **85**(2):159–169.

[176] Hu D and Cross JC 2010. Development and function of trophoblast giant cells in the rodent placenta. *The International Journal of Developmental Biology* **54**(2-3):341–354.

[177] Gupta S 2000. Hepatic polyploidy and liver growth control. *Seminars in cancer biology* **10**(3):161–171.

[178] Mollova M, Bersell K, Walsh S, Savla J, Das LT, et al. 2013. Cardiomyocyte proliferation contributes to heart growth in young humans. *Proceedings of the National Academy of Sciences* **110**(4):1446–1451.

[179] Ravid K, Lu J, Zimmet JM, and Jones MR 2002. Roads to polyploidy: the megakaryocyte example. *Journal of cellular physiology* **190**(1):7–20.

[180] Gandarillas A 2012. The mysterious human epidermal cell cycle, or an oncogene-induced differentiation checkpoint. *Cell cycle (Georgetown, Tex.)* **11**(24):4507–4516.

[181] McCrann DJ, Nguyen HG, Jones MR, and Ravid K 2008. Vascular smooth muscle cell polyploidy: an adaptive or maladaptive response? *Journal of cellular physiology* **215**(3):588–592.

[182] Kriz W, Hahnel B, Rosener S, and Elger M 1995. Long-term treatment of rats with FGF-2 results in focal segmental glomerulosclerosis. *Kidney international* **48**(5):1435–1450.

[183] Edgar BA, Zielke N, and Gutierrez C 2014. Endocycles: a recurrent evolutionary innovation for post-mitotic cell growth. *Nature Reviews Molecular Cell Biology* **15**(3):197–210.

[184] Gendreau E, Hofte H, Grandjean O, Brown S, and Traas J 1998. Phytochrome controls the number of endoreduplication cycles in the Arabidopsis thaliana hypocotyl. *The Plant journal : for cell and molecular biology* **13**(2):221–230.

[185] Sugimoto-Shirasu K and Roberts K 2003. “Big it up”: endoreduplication and cell-size control in plants. *Current Opinion in Plant Biology* **6**(6):544–553.

[186] Kondorosi E, Roudier F, and Gendreau E 2000. Plant cell-size control: growing by ploidy? *Current Opinion in Plant Biology* **3**(6):488–492.

[187] Melaragno JE, Mehrotra B, and Coleman AW 1993. Relationship between Endopolyploidy and Cell Size in Epidermal Tissue of Arabidopsis. *The Plant Cell* **5**(11):1661–1668.

[188] Tsumoto Y, Yoshizumi T, Kuroda H, Kawashima M, Ichikawa T, et al. 2006. Light-dependent polyploidy control by a CUE protein variant in Arabidopsis. *Plant molecular biology* **61**(4-5):817–828.

[189] Gendreau E, Traas J, Desnos T, Grandjean O, Caboche M, et al. 1997. Cellular basis of hypocotyl growth in Arabidopsis thaliana. *Plant physiology* **114**(1):295–305.

[190] Schrader A, Welter B, Hulskamp M, Hoecker U, and Uhrig JF 2013. MIDGET connects COP1-dependent development with endoreduplication in Arabidopsis thaliana. *The Plant journal : for cell and molecular biology* **75**(1):67–79.

[191] Lee HO, Davidson JM, and Duronio RJ 2009. Endoreplication: polyploidy with purpose. *Genes & Development* **23**(21):2461–2477.

[192] Donnelly PM, Bonetta D, Tsukaya H, Dengler RE, and Dengler NG 1999. Cell cycling and cell enlargement in developing leaves of Arabidopsis. *Developmental biology* **215**(2):407–419.

[193] Andriankaja M, Dhondt S, De Bodt S, Vanhaeren H, Coppens F, et al. 2012. Exit from proliferation during leaf development in Arabidopsis thaliana: a not-so-gradual process. *Developmental Cell* **22**(1):64–78.

[194] Kalve S, De Vos D, and Beemster GTS 2014. Leaf development: a cellular perspective. *Frontiers in plant science* **5**:362.

[195] Wagner GJ, Wang E, and Shepherd RW 2004. New approaches for studying and exploiting an old protuberance, the plant trichome. *Annals of botany* **93**(1):3–11.

[196] Ishida T, Kurata T, Okada K, and Wada T 2008. A genetic regulatory network in the development of trichomes and root hairs. *Annual Review of Plant Biology* **59**:365–386.

[197] Jakoby MJ, Falkenhan D, Mader MT, Brininstool G, Wischnitzki E, et al. 2008. Transcriptional profiling of mature Arabidopsis trichomes reveals that NOECK encodes the MIXTA-like transcriptional regulator MYB106. *Plant physiology* **148**(3):1583–1602.

[198] Lieckfeldt E, Simon-Rosin U, Kose F, Zoeller D, Schliep M, et al. 2008. Gene expression profiling of single epidermal, basal and trichome cells of Arabidopsis thaliana. *Journal of plant physiology* **165**(14):1530–1544.

[199] Marks MD, Wenger JP, Gilding E, Jilk R, and Dixon RA 2009. Transcriptome analysis of Arabidopsis wild-type and gl3-sst sim trichomes identifies four additional genes required for trichome development. *Molecular Plant* **2**(4):803–822.

[200] Folkers U, Berger J, and Hülskamp M 1997. Cell morphogenesis of trichomes in Arabidopsis: differential control of primary and secondary branching by branch initiation regulators and cell growth. *Development* **124**(19):3779–3786.

[201] Hulskamp M, Miséra S, and Jürgens G 1994. Genetic dissection of trichome cell development in Arabidopsis. *Cell* **76**(3):555–566.

[202] Schnittger A, Jurgens G, and Hülskamp M 1998. Tissue layer and organ specificity of trichome formation are regulated by GLABRA1 and TRIPTYCHON in Arabidopsis. *Development* **125**(12):2283–2289.

[203] Breuer C, Morohashi K, Kawamura A, Takahashi N, Ishida T, et al. 2012. Transcriptional repression of the APC/C activator CCS52A1 promotes active termination of cell growth. *The EMBO journal* **31**(24):4488–4501.

[204] Downes BP, Stupar RM, Gingerich DJ, and Vierstra RD 2003. The HECT ubiquitin-protein ligase (UPL) family in Arabidopsis: UPL3 has a specific role in trichome development. *The Plant journal : for cell and molecular biology* **35**(6):729–742.

[205] Breuer C, Braidwood L, and Sugimoto K 2014. Endocycling in the path of plant development. *Current Opinion in Plant Biology* **17**:78–85.

[206] Bramsiepe J, Wester K, Weinl C, Roodbarkelari F, Kasili R, et al. 2010. Endoreplication controls cell fate maintenance. *PLoS genetics* **6**(6):e1000996.

[207] Brininstool G, Kasili R, Simmons LA, Kirik V, Hulskamp M, et al. 2008. Constitutive Expressor Of Pathogenesis-related Genes5 affects cell wall biogenesis and trichome development. *BMC Plant Biology* **8**:58.

[208] De Veylder L, Larkin JC, and Schnittger A 2011. Molecular control and function of endoreplication in development and physiology. *Trends in Plant Science* **16**(11):624–634.

[209] Weinl C, Marquardt S, Kuijt SJH, Nowack MK, Jakoby MJ, et al. 2005. Novel functions of plant cyclin-dependent kinase inhibitors, ICK1/KRP1, can act non-cell-autonomously and inhibit entry into mitosis. *Plant Cell* **17**(6):1704–1722.

[210] Benfey PN, Linstead PJ, Roberts K, Schiefelbein JW, Hauser MT, et al. 1993. Root development in Arabidopsis: four mutants with dramatically altered root morphogenesis. *Development* **119**(1):57–70.

[211] Petricka JJ, Winter CM, and Benfey PN 2012. Control of Arabidopsis root development. *Annual Review of Plant Biology* **63**:563–590.

[212] Ishida T, Fujiwara S, Miura K, Stacey N, Yoshimura M, et al. 2009. SUMO E3 ligase HIGH PLOIDY2 regulates endocycle onset and meristem maintenance in Arabidopsis. *Plant Cell* **21**(8):2284–2297.

[213] Vanstraelen M, Baloban M, Da Ines O, Cultrone A, Lammens T, et al. 2009. APC/CCCS52A complexes control meristem maintenance in the Arabidopsis root. *Proceedings of the National Academy of Sciences* **106**(28):11806–11811.

[214] Wen B, Nieuwland J, and Murray JAH 2013. The Arabidopsis CDK inhibitor ICK3/KRP5 is rate limiting for primary root growth and promotes growth through cell elongation and endoreduplication. *Journal of experimental botany* **64**(4):1135–1144.

[215] Ishida T, Adachi S, Yoshimura M, Shimizu K, Umeda M, et al. 2010. Auxin modulates the transition from the mitotic cycle to the endocycle in Arabidopsis. *Development* **137**(1):63–71.

[216] Dello Ioio R, Linhares FS, Scacchi E, Casamitjana-Martinez E, Heidstra R, et al. 2007. Cytokinins determine Arabidopsis root-meristem size by controlling cell differentiation. *Current biology : CB* **17**(8):678–682.

[217] Dello Ioio R, Nakamura K, Moubayidin L, Perilli S, Taniguchi M, et al. 2008. A genetic framework for the control of cell division and differentiation in the root meristem. *Science* **322**(5906):1380–1384.

[218] Takahashi N, Kajihara T, Okamura C, Kim Y, Katagiri Y, et al. 2013. Cytokinins control endocycle onset by promoting the expression of an APC/C activator in Arabidopsis roots. *Current biology : CB* **23**(18):1812–1817.

[219] Parsons RF 2009. Hypocotyl hairs: an historical perspective. *Australian Journal of Botany* **57**(2):106–108.

[220] Sliwinska E, Bassel GW, and Bewley JD 2009. Germination of Arabidopsis thaliana seeds is not completed as a result of elongation of the radicle but of the adjacent transition zone and lower hypocotyl. *Journal of experimental botany* **60**(12):3587–3594.

[221] Sliwinska E, Mathur J, and Bewley JD 2012. Synchronously developing collet hairs in Arabidopsis thaliana provide an easily accessible system for studying nuclear movement and endoreduplication. *Journal of experimental botany* **63**(11):4165–4178.

[222] Tominaga-Wada R, Ishida T, and Wada T 2011. New insights into the mechanism of development of Arabidopsis root hairs and trichomes. *Int Rev Cell Mol Biol* **286**:67–106.

[223] Yi K, Menand B, Bell E, and Dolan L 2010. A basic helix-loop-helix transcription factor controls cell growth and size in root hairs. *Nature Genetics* **42**(3):264–267.

[224] Cookson SJ and Granier C 2006. A dynamic analysis of the shade-induced plasticity in Arabidopsis thaliana rosette leaf development reveals new components of the shade-adaptative response. *Annals of botany* **97**(3):443–452.

[225] Adachi S, Minamisawa K, Okushima Y, Inagaki S, Yoshiyama K, et al. 2011. Programmed induction of endoreduplication by DNA double-strand breaks in Arabidopsis. *Proceedings of the National Academy of Sciences* **108**(24):10004–10009.

[226] Claeys H, Skirycz A, Maleux K, and Inzé D 2012. DELLA signaling mediates stress-induced cell differentiation in Arabidopsis leaves through modulation of anaphase-promoting complex/cyclosome activity. *Plant physiology* **159**(2):739–747.

[227] Breuer C, Ishida T, and Sugimoto K 2010. Developmental control of endocycles and cell growth in plants. *Current Opinion in Plant Biology* **13**(6):654–660.

[228] Ste D 2003. Effect of chilling on DNA endoreplication in root cortex cells and root hairs of soybean seedlings. *Biologia plantarum* **47**(3):333–339.

[229] Engelen-Eigles G, Jones RJ, and Phillips RL 2001. DNA endoreduplication in maize endosperm cells is reduced by high temperature during the mitotic phase. *Crop science* **41**(4):1114–1121.

[230] Bertin N 2005. Analysis of the tomato fruit growth response to temperature and plant fruit load in relation to cell division, cell expansion and DNA endoreduplication. *Annals of botany* **95**(3):439–447.

[231] Lee HC, Chen YJ, Markhart AH, and Lin TY 2007. Temperature effects on systemic endoreduplication in orchid during floral development. *Plant Science* **172**(3):588–595.

[232] Jovtchev G, Barow M, Meister A, and Schubert I 2007. Impact of environmental and endogenous factors on endopolyploidization in angiosperms. *Environmental and Experimental Botany* **60**(3):404–411.

[233] Chandran D, Inada N, Hather G, Kleindt CK, and Wildermuth MC 2010. Laser microdissection of Arabidopsis cells at the powdery mildew infection site reveals site-specific processes and regulators. *Proceedings of the National Academy of Sciences* **107**(1):460–465.

[234] de Almeida Engler J, Engler G, and Gheysen G 2011. *Unravelling the Plant Cell Cycle in Nematode Induced Feeding Sites*. Genomics and molecular genetics of plant-nematode interactions. Springer Netherlands.

[235] Vieira P, Kyndt T, Gheysen G, and Engler JdA 2013. An insight into critical endocycle genes for plant-parasitic nematode feeding sites establishment. *Plant signaling & behavior* **8**(1559-2316):e24223.

[236] Hamdoun S, Liu Z, Gill M, Yao N, and Lu H 2013. Dynamics of Defense Responses and Cell Fate Change during Arabidopsis-Pseudomonas syringae Interactions. *PloS one* **8**(12):e83219.

[237] Lingua G, Fusconi A, and Berta G 2001. The nucleus of differentiated root plant cells: modifications induced by arbuscular mycorrhizal fungi. *European Journal of Histochemistry* **45**(1):9–20.

[238] Penterman J, Abo RP, De Nisco NJ, Arnold MFF, Longhi R, et al. 2014. Host plant peptides elicit a transcriptional response to control the Sinorhizobium meliloti cell cycle during symbiosis. *Proceedings of the National Academy of Sciences* **111**(9):3561–3566.

[239] BAINARD LD, BAINARD JD, NEWMASTER SG, and KLIRONOMOS JN 2011. Mycorrhizal symbiosis stimulates endoreduplication in angiosperms. *Plant, Cell & Environment* **34**(9):1577–1585.

[240] Barow M and Meister A 2003. Endopolyploidy in seed plants is differently correlated to systematics, organ, life strategy and genome size. *Plant, Cell & Environment* **26**(4):571–584.

[241] Yamasaki S, Shimada E, Kuwano T, Kawano T, and Noguchi N 2010. Continuous UV-B irradiation induces endoreduplication and peroxidase activity in epidermal cells surrounding trichomes on cucumber cotyledons. *Journal of radiation research* **51**(2):187–196.

[242] Gendreau E, Orbovic V, Höfte H, and Traas J 1999. Gibberellin and ethylene control endoreduplication levels in the Arabidopsis thaliana hypocotyl. *Planta* **209**(4):513–516.

[243] Perazza DD, Herzog MM, Hülskamp MM, Brown SS, Dorne AMA, et al. 1999. Trichome cell growth in Arabidopsis thaliana can be derepressed by mutations in at least five genes. *Genetics* **152**(1):461–476.

[244] Vinod PK, Freire P, Rattani A, Ciliberto A, Uhlmann F, et al. 2011. Computational modelling of mitotic exit in budding yeast: the role of separase and Cdc14 endocycles. *Journal of The Royal Society Interface* **8**(61):1128–1141.

[245] Queralt E, Lehane C, Novak B, and Uhlmann F 2006. Downregulation of PP2ACdc55 Phosphatase by Separase Initiates Mitotic Exit in Budding Yeast. *Cell* **125**(4):719–732.

[246] Tóth A, Queralt E, Uhlmann F, and Novak B. Mitotic exit in two dimensions. *Journal of theoretical biology* **248**(3):560–573.

[247] Zielke N, Kim KJ, Tran V, Shibutani ST, Bravo MJ, et al. 2011. Control of Drosophila endocycles by E2F and CRL4(CDT2). *Nature* **480**(7375):123–127.

[248] Roodbarkelari F, Bramsiepe J, Weinl C, Marquardt S, Novak B, et al. 2010. Cullin 4-ring finger-ligase plays a key role in the control of endoreplication cycles in Arabidopsis trichomes. *Proceedings of the National Academy of Sciences of the United States of America* **107**(34):15275–15280.

[249] Traas J, Hulskamp M, Gendreau E, and Hofte H 1998. Endoreduplication and development: rule without dividing? *Current Opinion in Plant Biology* **1**(6):498–503.

[250] Kondorosi E and Kondorosi A 2004. Endoreduplication and activation of the anaphase-promoting complex during symbiotic cell development. *FEBS Letters* **567**(1):152–157.

[251] Hayashi K, Hasegawa J, and Matsunaga S 2013. The boundary of the meristematic and elongation zones in roots: endoreduplication precedes rapid cell expansion. *Scientific reports* **3**:2723.

[252] Cookson SJ, Radziejwoski A, and Granier C 2006. Cell and leaf size plasticity in Arabidopsis: what is the role of endoreduplication? *Plant, Cell & Environment* **29**(7):1273–1283.

[253] Gegas VC, Wargent JJ, Pesquet E, Granqvist E, Paul ND, et al. 2014. Endopolyploidy as a potential alternative adaptive strategy for Arabidopsis leaf size variation in response to UV-B. *Journal of experimental botany* **65**(10):2757–2766.

[254] Heidstra R, Welch D, and Scheres B 2004. Mosaic analyses using marked activation and deletion clones dissect Arabidopsis SCARECROW action in asymmetric cell division. *Genes & Development* **18**(16):1964–1969.

[255] Cartwright DA, Brady SM, Orlando DA, Sturmfels B, and Benfey PN 2009. Reconstructing spatiotemporal gene expression data from partial observations. *Bioinformatics* **25**(19):2581–2587.

[256] Beeckman T and De Smet I 2014. Pericycle. *Current biology : CB* **24**(10):R378–9.

[257] Yi D, Alvim Kamei CL, Cools T, Vanderauwera S, Takahashi N, et al. 2014. The Arabidopsis SIAMESE-RELATED Cyclin-Dependent Kinase Inhibitors SMR5 and SMR7 Regulate the DNA Damage Checkpoint in Response to Reactive Oxygen Species. *The Plant Cell* **26**(1):296–309.

[258] Myers PN, Setter TL, Madison JT, and Thompson JF 1990. Abscisic Acid Inhibition of Endosperm Cell Division in Cultured Maize Kernels. *Plant physiology* .

[259] Artlip TS, Madison JT, and Setter TL 1995. Water deficit in developing endosperm of maize: cell division and nuclear DMA endoreduplication. *Plant, Cell & Environment* **18**(9):1034–1040.

[260] Valente P, Tao W, and Verbelen JP 1998. Auxins and cytokinins control DNA endoreduplication and deduplication in single cells of tobacco. *Plant Science* **134**(2):207–215.

[261] Dinneny JR, Long TA, Wang JY, Jung JW, Mace D, et al. 2008. Cell Identity Mediates the Response of Arabidopsis Roots to Abiotic Stress. *Science* **320**(5878):942–945.

[262] Grime JP and Mowforth MA 1982. Variation in genome size—an ecological interpretation .

[263] Murashige T and Skoog F 1962. A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. *Physiologia Plantarum* **15**(3):473–497.

[264] Iyer-Pascuzzi AS, Jackson T, Cui H, Petricka JJ, Busch W, et al. 2011. Cell identity regulators link development and stress responses in the Arabidopsis root. *Developmental Cell* **21**(4):770–782.

[265] Lewis DR, Olex AL, Lundy SR, Turkett WH, Fetrow JS, et al. 2013. A kinetic analysis of the auxin transcriptome reveals cell wall remodeling proteins that modulate lateral root development in Arabidopsis. *The Plant Cell* **25**(9):3329–3346.

[266] Zhang C, Gong FC, Lambert GM, and Galbraith DW 2005. Cell type-specific characterization of nuclear DNA contents within complex tissues and organs. *Plant methods* **1**(1):7.

[267] Zhang C, Barthelson RA, Lambert GM, and Galbraith DW 2008. Global characterization of cell-specific gene expression through fluorescence-activated sorting of nuclei. *Plant physiology* **147**(1):30–40.

[268] Hochberg Y and Benjamini Y 1990. More powerful procedures for multiple significance testing. *Statistics in Medicine* **9**(7):811–818.

[269] De Bodt S, Hollunder J, Nelissen H, Meulemeester N, and Inzé D 2012. CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations. *The New phytologist* **195**(3):707–720.

[270] Edgar R, Domrachev M, and Lash AE 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**(1):207–210.